Committee

Prof. Dr. Gerda Claeskens (Advisor)	*Katholieke Universiteit Leuven*
Prof. Dr. Christophe Croux (Advisor)	*Katholieke Universiteit Leuven*
Prof. Dr. Eva Cantoni	*University of Geneva*
Prof. Dr. Irène Gijbels	*Katholieke Universiteit Leuven*
Prof. Dr. Martina Vandebroek	*Katholieke Universiteit Leuven*

Acknowledgements

"The journey of a thousand miles begins with a single step"
Laozi, ancient Chinese philosopher

This thesis can be seen as the story of this "journey of a thousand miles", a journey which was replete with pleasant surprises, and very annoying setbacks. However, it omits one very important aspect of such a journey: the companions you start out with, and the friends you make along this road, all of them helping you make that journey, whether directly or indirectly. As such, a very heartfelt thank you and acknowledgement of their efforts is certainly warranted here.

The persons I would like to thank the most, if it is even possible to say who deserve the most gratitude, are my mother, may she rest in peace, and my father. They have been there for me always, encouraged me and guided me where needed, and generally, have been, and still are, great parents. My only deep regret is that she didn't live to see me complete my PhD, but I know she would be proud.

Also eligible for a great deal of gratitude are Prof. Dr. Gerda Claeskens and Prof. Dr. Christophe Croux, my two advisors. I realise it must have been far from an easy task to put up with my oddities for all this time, and especially keep my mind focussed and where it belongs, but they managed to do just that admirably well. Especially Prof. Dr. Croux deserves my gratitude for offering me this research opportunity in the first place, since at that time I was literally like a ship adrift, jobwise. By extension, I would also like to thank the other members of my doctoral committee, Prof. Dr. Eva Cantoni, Prof. Dr. Irène Gijbels, and Prof. Dr. Martina Vandebroek, for agreeing to seat in the committee, and especially for their insightful comments and pointed questions.

Next up are the great colleagues I had the pleasure of getting to know, and of having several discussions with, both serious and silly. So, ladies and gentlemen, thank you for the nice hours of conversation. Colleagues who I think should deserve a special prize are definitely the ones who I shared an office with over these years. Mario Pandelaere, Robrecht Van Goolen, Aurélie Lemmens, Kris Boudt, Fabrice Talla Nobibon, Fabrizio Consentino, and Derya Çalışkan, thank you all for being great officemates.

Quite important, but sometimes forgotten, are the reviewers of the papers I have submitted. Their careful and thorough reading of these papers, and their many insightful comments have certainly helped to improve the quality of the work presented here. For this, they deserve my gratitude. Thank you.

In the beginning of the acknowledgement I have expressed my gratitude for the two persons who brought me into this world, but let's certainly not forget the rest of the family: the grandparents, uncles and aunts, cousins, nephews and nieces. Thank you all.

What is a man without his friends? Therefore I extend my special gratitude towards my friends and acquaintances, and especially to the ladies and gentlemen of the Mechelse Budo Sporten Aikido, for so graciously being a good source of stress relief.

And finally, there are two very special mentions that I just have to give. The first special mention is for Mr. Noël Tourlouse, my mathematics teacher in my two last years at the Koninklijk Atheneum Pitzemburg. He really brought my enjoyment of the wonderful world of mathematics to a new level, and I owe him a lot for this. The second special mention is for Prof. Dr. Uwe Einmahl, my professor of Probability Theory during my years studying Mathematics at the Vrije Universiteit Brussel. He was the advisor of my undergraduate/master thesis (Dutch: licentiaatsverhandeling), and taught me the first steps of how to do research.

To everyone I mentioned, and the people important to me and this work whom I failed to mention above, thank you for all your support.

Johan Van Kerckhoven Leuven, December 2007

Summary

These days, many studies in the economical, medical, biochemical, and many other fields of research result in enormous amounts of data on a phenomenon. Examples of such datasets are customer data assessing credit risk (for banks), accident risk (for insurance companies), epidemiological studies, and genetic relevance studies. It also becomes more common that these datasets encompass a large number of variables, most of which are likely to be irrelevant in relation to the phenomenon under investigation. Therefore, methods are needed which select a group of variables, preferably as small as possible, or a proposed model, as sparse as possible, which still provides a sufficiently good model for the investigated phenomenon.

To this end, many different model-selection criteria have already been proposed, such as Akaike's Information Criterion (AIC), Bayesian or Schwarz' Information Criterion (BIC/SIC), Mallows' Cp , and more recently, the Focused Information Criterion (FIC). The first three of these criteria will allow the user to select one specific model to explain the phenomenon under investigation, irrespective of the later use of the model. While these criteria usually select a model with good overall performance, it might not be optimised for the proposed task, such as prediction for example. The latter criterion, FIC, overcomes this criticism and selects a model specifically suited for the task at hand, such that the selected model possibly has a better performance for that particular task.

In the first chapter of the thesis, we have considered the issue of prediction focussed variable selection in logistic regression models. In this particular setting, the FIC will select different models depending on the observation about which the

prediction is made, leading to more accurate predictions. This is of particular interest for business managers, who want to predict as accurately as possible whether a particular business venture will succeed or not. Other applications are for example in the medical field, where it is vital that patients are correctly diagnosed as having a certain disease or not.

The standard version of the FIC estimates the Mean Squared Error (MSE) of the estimator of the quantity of interest, here chosen to be the score of the observation of interest. In this chapter we have proposed more general versions of the FIC, allowing other risk measures such as one based on Lp-error. More importantly however, we have constructed a FIC using the misclassification probability as natural risk measure, since the goal is to accurately predict the binary outcome. The advantages of using an information criterion for selecting suitable models for prediction which depends on the new observation and on the selected risk measure have been illustrated by means of a simulation study and an application to a study on diabetic retinopathy.

In the second chapter of the thesis, we have applied FIC to select the autoregressive (AR) order of a stationary time series. Autoregressive time series are often used in economics to model a phenomenon, such as exchange rates or unemployment, over time. These models are then used to predict the value of that phenomenon for the near future. Especially for macro-economic phenomena, these predictions should be as accurate as possible, such that policy makers can rely on these predictions to make good decisions.

Originally, the focussed information criterion has been proposed for a fixed model set, where the largest model under consideration doesn't vary when observations are added. In this chapter, we have provided a theoretical foundation such that the FIC can be applied when the maximal AR-order under consideration increases towards infinity as the length of the time series increases. This result is needed for two reasons. First of all, the number of variables to select from is in principle infinity in the setting of autoregressive models. More importantly however, we wished to examine the asymptotic efficiency properties of the FIC and compare it to AIC for model-order selection. This investigation has been conducted by means of a detailed simulation study, studying both the

special two-series setting where AIC will asymptotically select the most accurate model for prediction (lowest MSE), as in the much more common single series setting, where AIC has the same property. In this study, we have observed that the performance of the models selected by FIC is very close to that of the models selected by AIC, and that the difference in performance becomes smaller as the length of the series increases.

The FIC can also be used to select the best models for estimating the impulse response function of a series at a certain lag. In this case, the relative performance of FIC with respect to AIC varies with the parameters of the true data-generating model, and neither uniformly dominates the other. Finally, we have illustrated that the FIC can be applied easily towards more complicated variable selection tasks in the time series framework, such as simultaneous selection of both regression variables and the autoregressive order of the error terms.

The criteria outlined in the paragraphs above have one major drawback however. As these are likelihood-based information criteria, they are of little to no use when the number of variables increases beyond the number of available observations. First, we will need an alternative to maximum likelihood estimation to actually be able to estimate the model parameters. Support Vector Machines provide a means to do classification when the number of variables (greatly) exceeds the number of available observations. Nevertheless, it is still recommended to reduce the dimension of your input space to increase the predictive performance of the estimated model. Several techniques have already been proposed to perform variable selection in this setting, though few of them rely on information criteria. Methods which rely on such criteria are for example cross-validated error rate based criteria, or the Kernel Regularisation Information Criterion (KRIC).

In the third chapter of the thesis, we have developed two new information criteria (SVMICa and SVMICb) which can be used for variable selection in the SVM setting. These newly proposed criteria have the advantage that they incur less computational overhead than the already existing criteria, and as such, are faster to evaluate. Secondly, we have linked SVMICa to the aforementioned KRIC, as an approximation under certain conditions. We have then performed an extensive simulation study in which we examined the properties of SVMICa/b,

and we found that the models selected by these criteria have decent predictive power. Moreover, the simulation study indicated that SVMICb exhibits the property of asymptotic consistency. Finally, a test on real data verified the adequate performance of the newly developed criteria.

A different issue, but one which is still very important in predictive modelling, is how efficient an estimation method for a certain model is. Generally speaking, there is a trade-off between the efficiency of an estimation method, and how generally applicable or robust that method is. Hence, examining the efficiencies of those estimation methods provides the answer to the question of what price the researcher pays (in terms of efficiency) for the additional generality and/or robustness of the used estimation method.

In the final chapter of this thesis, we have examined the classification efficiencies of a group of decision rules which are known as Convex Risk Minimisation (CRM) rules. These methods are a very flexible class of estimation methods for the decision function for binary classification, in the sense that they can easily be used for estimating non-linear decision functions. We have compared this class of rules against the well-known Fisher's linear discriminant rule, and this in the setting of two normally distributed populations with equal variance, where it is known that Fisher's rule is efficient. To compute the classification efficiencies, we have used influence functions. First of all, we have obtained a general expression for the influence function of a Fisher-consistent CRM technique, in the sense that the decision rule achieves the minimal obtainable generalisation error. We have also obtained sufficient conditions for such Convex Risk Minimisation rules to be Fisher-consistent. Then, we have performed a detailed case-by-case analysis for a number of CRM methods, and we have found that reasonably balanced populations which are badly separated, the CRM still have decent efficiency, above 50%, while being much more flexible than the efficient Fisher's rule.

Samenvatting

Op de dag van vandaag worden er enorm veel gegevens verzameld in studies over economische, medische, biochemische en vele andere fenomenen. Voorbeelden van zulke datasets zijn bijvoorbeeld gegevens over klanten voor het bepalen van hun kredietrisico (voor banken), hun risico op ongevallen (voor verzekeringsmaatschappijen). Andere voorbeelden zijn onder andere epidemiologische studies, en studies naar genetische relevantie. Ook gebeurt het steeds meer dat deze datasets veel verschillende variabelen bevatten, waarvan de meeste waarschijnlijk niets te maken hebben met het onderzochte fenomeen. Daarom zijn er technieken nodig die een groep van variabelen kunnen selecteren, liefst zo klein mogelijk, of een zo eenvoudig mogelijk model, dat toch een goed model is voor het onderzochte fenomeen.

Daartoe zijn er al verschillende modelselectiecriteria ontwikkeld, zoals Akaike's Informatiecriterium (AIC), het Bayesiaans of Schwarz' Informatiecriterium (BIC/ SIC), het C_p criterium van Mallows, and meer recent, het Focussed Informatiecriterium (FIC). De eerste drie criteria in deze lijst laten toe van één bepaald model te kiezen om het onderzochte fenomeen te verklaren, waarvoor dit model ook gebruikt zal worden. Hoewel deze criteria doorgaans een model kiezen dat behoorlijk werkt, is het niet noodzakelijk optimaal voor het uiteindelijke doel, bijvoorbeeld om voorspellingen te maken. Het laatste criterium echter, het FIC, heeft dit probleem niet en zal een model kiezen dat op maat gemaakt is voor wat de onderzoeker voor ogen heeft, waardoor het gekozen model mogelijk beter presteert voor dat bepaald doel.

In het eerste hoofdstuk van deze thesis bekijken we het probleem van doel-

gerichte variabelenselectie in het logistisch regressiemodel. Hier zal het FIC verschillende modellen kiezen naargelang de observatie waarover de voorspelling wordt gemaakt, wat tot nauwkeurigere voorspellingen zal leiden. Dit is vooral interessant voor zakenmanagers als ze willen voorspellen dat een bepaalde investering zal renderen of niet. Een andere toepassing bevindt zich in de medische wereld, waar het van levensbelang is dat patiënten een correcte diagnose krijgen dat ze al dan niet een bepaalde ziekte hebben.

De gewone FIC schat de gemiddelde kwadratische fout van de schatter van de parameter die ons interesseert, waarbij we hier de score van de te voorspellen observatie kiezen. In dit hoofdstuk hebben we een algemenere versie van FIC voorgesteld met een algemene risicomaat gebaseerd op de L_p-fout. De hoofdverwezenlijking hier is het opstellen van een FIC waarbij de kans op een foute voorspelling als risicomaat wordt gebruikt, vermits we een ja/nee uitkomst willen voorspellen. De voordelen van het gebruik van een informatiecriterium dat zijn model kiest afhankelijk van de te voorspellen observatie worden aangetoond aan de hand van een simulatiestudie en een toepassing op een medische studie.

In het tweede hoofdstuk van de thesis passen we het FIC toe op het kiezen van de autoregressie (AR) orde van een stationaire tijdreeks. Autoregressieve tijdreeksen worden in economie vaak gebruikt om een fenomeen zoals wisselkoersen of werkloosheidsgraad over de tijd te modelleren. Deze modellen worden dan gebruikt om dit fenomeen te voorspellen voor de (nabije) toekomst. Deze voorspellingen moeten zo nauwkeurig mogelijk zijn, dit in het bijzonder voor macroeconomische fenomenen, zodanig dat de beleidsmensen hierop kunnen vertrouwen voor het nemen van goede beslissingen.

Het focussed informatiecriterium was oorspronkelijk gedefinieerd voor een vaste groep van modellen, waarbij het grootste beschouwd model niet verandert als er observaties bijkomen. In dit hoofdstuk ontwikkelden we het FIC verder zodanig dat dit criterium kan gebruikt worden in de situatie waar de maximale AR orde van de beschouwde modellen naar oneindig gaat als de lengte van de tijdreeks stijgt. We hebben dit resultaat voor twee redenen nodig. Eerst en vooral is het aantal mogelijke variabelen theoretisch oneindig als we werken met autoregressieve modellen. Een belangrijkere reden is dat we de asymptotische ef-

ficiëntie van FIC wensen te onderzoeken, en dit willen vergelijken met AIC voor modelorde selectie. We hebben dit onderzocht aan de hand van een uitgebreide simulatiestudie, waarbij we zowel het geval van twee tijdreeksen hebben onderzocht, waar AIC asymptotisch de meest nauwkeurige modellen selecteert, als het geval van één enkele tijdreeks, waar AIC deze eigenschap ook heeft. Gedurende deze studie hebben we gemerkt dat de prestaties van de modellen geselecteerd door FIC zeer dicht liggen bij de prestaties van de modellen geselecteerd door AIC en dat dit verschil kleiner wordt als de lengte van de tijreeks stijgt.

Het FIC kan ook gebruikt worden om het beste model te kiezen voor het schatten van de impulsresponsfunctie voor een bepaalde lag. In dit geval zien we dat de prestaties van FIC en AIC sterk variëren naargelang de parameters van het echte, datagenererend model veranderen, en dat geen van beide uniform beter is dan het andere. Ook hebben we aangetoond dat FIC eenvoudig kan worden toegepast voor moeilijkere variabelenselectie problemen voor tijdreeksen, zoals het tegelijkertijd selecteren van de regressievariabelen en de AR orde van de residuen.

De criteria in de voorgaande paragrafen hebben één groot nadeel. Omdat ze gebaseerd zijn op de likelihood van de gegevens, kunnen ze niet gebruikt worden als het aantal variabelen groter is dan het aantal observaties. Daarom hebben we eerst een alternatief voor maximum likelihood schatters nodig, zodanig dat we de parameters van het model kunnen schatten. De Support Vector Machine (SVM) laat binaire classificatie toe als het aantal variabelen (veel) groter is dan het aantal observaties. Het is echter nog altijd aan te raden om de dimensie van de ruimte van de observaties te verkleinen, omdat dit de voorspellende prestaties van het model kan vergroten. Er zijn reeds verschillende technieken om variabelenselectie te doen voor de SVM, maar weinigen ervan werken met informatiecriteria. Technieken die toch op criteria zijn gebaseerd zijn bijvoorbeeld deze gebaseerd op de crossvalidatie voorspellingsfout, of het Kernel Regularisatie Informatiecriterium (KRIC).

In het derde hoofdstuk van deze thesis hebben we twee nieuwe informatiecriteria ontwikkeld (SVMICa en SVMICb) die voor variabelenselectie in SVM's kunnen worden gebruikt. Deze nieuwe criteria hebben als voordeel dat ze niet

zo veel extra berekeningen vragen als de bestaande criteria, en dat ze dus sneller te berekenen zijn. Ook hebben we het SVMICa gekoppeld aan het KRIC, als een benadering onder bepaalde voorwaarden. Daarna hebben we een uitgebreide simulatiestudie uitgevoerd waarin we de eigenschappen van SVMICa/b hebben onderzocht, en we hebben gezien dat de modellen geselecteerd door deze criteria degelijke voorspellende eigenschappen hebben. Daarenboven blijkt SVMICb de asymptotische consistentie eigenschap te hebben. Deze goede eigenschappen werden ook bevestigd gedurende een test op een aantal echte datasets.

Een andere kwestie die toch zeer belangrijk is in het voorspellend modelleren is, is de vraag hoe efficiënt een schattingsmethode voor een bepaald model is. Doorgaans moet je een keuze maken tussen efficiëntie van de methode, en hoe algemeen toepasbaar of hoe robuust die methode is. Het onderzoeken van deze efficiënties laat ons dus toe te zien welke prijs (in termen van efficiëntie) je betaalt voor het gebruik van algemenere en/of robuustere schattingsmethoden.

In het laatste hoofdstuk van de thesis hebben we de classificatie-efficiëntie van een groep beslissingsregels, gekend als de Convex Risico Minimalisatie (CRM) regels, onderzocht. Deze methoden zijn een zeer flexibele groep van schattingstechnieken voor het schatten van de beslissingfunctie in binaire classificatie, in de zin dat deze eenvoudig kunnen aangewend worden voor niet-lineaire problemen. We hebben de CRM technieken vergeleken met de bekende lineaire discriminatieregel van Fisher, dit in het geval van twee normaalverdeelde populaties met gelijke variantie. In deze situatie weten we dat de regel van Fisher efficiënt is. Om die classificatie-efficiënties te bereken, maken we gebruik van invloedsfuncties. Eerst en vooral hebben we een theoretische uitdrukking gevonden voor deze invloedsfuncties voor Fisher-consistente CRM regels, regels die de laagst mogelijke voorspellingsfout hebben. Ook hebben we voldoende condities opgesteld waarvoor zulke Convex Risico Minimalisatie methodes Fisher-consistent zijn. Daarna hebben we een gedetailleerde analyse gedaan voor een aantal CRM methodes, en we hebben gevonden dat voor redelijk gebalanceerde, slecht scheidbare populaties, de CRM methodes redelijk efficiënt zijn, met efficiëntie boven de 50%, terwijl ze toch veel flexibeler zijn dan de efficiënte regel van Fisher.

Contents

Chapter 1

Introduction

The papers bundled in this dissertation cover two topics which are of considerable interest in statistical modelling. The first part, consisting of three essays, deal with the issue of selecting the "best" subset of variables from your data, where we define "best" in the sense of variables with good predictive power. In the second part, consisting of the last essay, we stepped away from the variable selection question and we investigated how well a statistical model actually performs, how efficient it is in terms of predictive performance.

Since antiquity, one of mankind's drives has been to explain a certain phenomenon, and link it to several others. This can range from something as simple as knowing that the position of the sun in the sky depends on what time of day it is (or vice versa), over the various laws in physics, to finding the relation between a person's income when he or she was 30 years old, and various pieces of information about that person, such as the birth date, gender, education level, you name it. Although there are probably millions of other examples of relations between variables, they all follow the same pattern. Suppose that Y is a certain phenomenon which can be measured and which interests you, whether this is the average fuel consumption of a car in miles-per-gallon, the price of MegaHuge, Inc. stocks, or whether someone is employed or not. Also assume that X is a list, or vector, of possibly relevant information, such as weight and engine power

of the car for the car example, the reported profits and the general state of the economy for the MegaHuge, Inc. example, or the person's education level, and the general conditions of the job market in the employment example. Then the vector X contains the explanatory, or predictor variables, and Y the response variable. The relation between the explicative variables X and the response variable Y can be written as

$$Y = f(X) + \varepsilon, \tag{1.1}$$

where $f(\cdot)$ can be any function of the predictors, and where ε is a term which stands for all the possible errors which can appear, such as measurement errors, or variables which we have not observed. The equation (1.1) is called a *statistical or stochastic model of Y with respect to X*.

In *statistical modelling* the researcher wants to find a model which can accurately describe the relation between the explicative variables X and the response variable Y. For this, he starts from a *(training) sample* consisting of n observations of (X, Y), which we denote (x_i, y_i) with $i = 1, \ldots, n$. Then, he *estimates* a function $\hat{f}$, such that the predicted responses $\hat{y}_i = \hat{f}(x_i)$ are "close" to the true responses y_i. In other words, that the errors, or *residuals* $r_i = \hat{y}_i - y_i$ are small. One commonly used measure of the (in-)accuracy of a model (1.1) is the *Mean Squared Error* (MSE):

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} r_i^2, \tag{1.2}$$

the average of the squares of the residuals. The lower this value is, the more accurate the predicted model.

Naturally, trying to estimate a completely arbitrary function f is next to impossible. Therefor the researcher makes several assumptions about this function. The most common assumptions are explicit assumptions about the shape, or the form of the function. A popular and well-studied example is the *linear (regression) model*, where we assume that

$$y = \alpha + \beta^t X + \varepsilon. \tag{1.3}$$

In this model, α is called the *intercept*, and the elements of the vector β are called

the *slope parameters*. The fuel usage example and the MegaHuge Inc. example mentioned above can be modelled with a linear regression model.

A special case of a linear regression model arises when you observe a certain phenomenon, such as the Euro/US Dollar exchange rate, over a period of time. We denote this *time series* as z_t, where $t = 1, \ldots, T$. Continuing our example, assume that we want to find the relation between today's exchange rate, and the exchange rates of the past three days. For this, we estimate the model

$$z_t = \phi_1 z_{t-1} + \phi_2 z_{t-2} + \phi_3 z_{t-3} + \varepsilon,$$

which we call an *autoregressive model of order 3*. This is a special case of the linear model described in (1.3), with Z_t as the response variables, and Z_{t-1}, Z_{t-2}, and Z_{t-3} (the *lagged* series of lag 1, 2, and 3 respectively) the explicative variables.

We use a different kind of model when the response variable Y can take only two values: 0, a certain event is not observed, and 1, a certain event is observed. We call Y a binary response variable, and the model relating Y to the explicative variables X is called a *binary choice model*. Generally, the model is defined by a *decision function* f of the predictor variables, such that

$$\begin{aligned} Y &= 1 \quad \text{for } f(X) > 0, \text{ and} \\ Y &= 0 \quad \text{otherwise.} \end{aligned} \tag{1.4}$$

Generally, given a sample of n observations (x_i, y_i), the above condition will not be satisfied for all observations. Hence, the modelling step consists here of finding a function $\hat{f}$ such that the *decision rule* (1.4) is violated "as little as possible".

One popular binary choice model is the *logistic regression model*. In this model, the decision function is of the form

$$f(X) = \alpha + \beta^t X,$$

and we estimate the parameters α and β from the model

$$P(Y = 1|X) = \frac{\exp(\alpha + \beta^t X)}{1 + \exp(\alpha + \beta^t X)} \stackrel{\text{def}}{=} F(\alpha + \beta^t X). \tag{1.5}$$

We call the *inverse link function* $F(\cdot)$ the *inverse logit function*.

Another class of binary choice models is the class of Convex Risk Minimisation techniques. This is a highly flexible class of methods which estimates the decision function f by minimising the average *loss*

$$\frac{1}{n}\sum_{i=1}^{n} L(y_i f(x_i)),$$

where $L(\cdot)$ is the *loss function*, a function which is positive, continuous, convex, and decreasing. Examples of such loss function are the hinge loss $L(u) = [1 - u]_+ = \max(0, 1 - u)$ which leads to the Support Vector Machine, and $L(u) = \log(1 + \exp(-u))$ which is the loss function for (Kernel) Logistic Regression.

As mentioned above, one of the goals of statistical modelling is to find an accurate model for a dataset (x_i, y_i). This is especially important if that model will be used later on to make predictions for new observations x_0. One of the most obvious ways of making a model more accurate is to give more information to the model, in other words, to use more variables. This is not a problem, especially in current times where enormous amounts of information can be gathered and stored quite easily. However, estimating a model with a high number of explicative variables, especially if many of them add little or no information, has several drawbacks. First, models with lots of variables take longer to estimate. This is not a very serious problem since computational power and speed increases rapidly. The second drawback is a more serious one. Models with lots of variables can become very complex and as such, the researcher is hard-pressed to interpret them. In essence, it means that the model which was supposed to clarify the relation between the response variable Y and the predictor variables X, end up making the relation even more obscure! The third drawback, and the most serious one, is that estimates of models with a high number of variables are more sensitive to the data than models with a low number of variables. This means that a slight change in the data can cause a large change in the estimated model. As such, predictions made with such a rich model are not necessarily as reliable as predictions made with a more compact model. Finally, when a model includes many variables, the variance of the estimated parameters increases when compared to a model with less parameters. Similarly to the previous reason, this

will also result in a greater variance for predictions made with the larger model.

To keep the number of variables low, the researcher can use her expertise, or call upon someone else's expertise, to manually pick out the variables which she considers to be important, but this is rarely a practical solution. Indeed, manually selecting the variables might be too time-consuming, and the selection is quite subjective. Also, the needed expertise might be very expensive, or even worse, simply not available! For this reason, various methods have been developed to select the important variables in a dataset, based on the data itself.

A popular method for comparing models using different subsets of variables of the same dataset is by attaching an *information criterion* to the model. This is in essence a numeric value indicating how well (or how bad) that particular model explains that particular dataset, with a penalisation added for the complexity of the model. The most popular criteria of this type are of the form

$$-2 \log L(x_i, y_i) + C(n)p.$$

Here, $L(x_i, y_i)$ is the *likelihood* of the model given the data (see Pawitan, 2001, for more details), p is the number of variables used in that particular model, and $C(n)$ is a positive penalty function, possibly depending on the number of observations n. Well known criteria of this form are Akaike's Information Criterion (Akaike, 1974), where $C(n) = 2$, and Bayes' Information Criterion (Schwarz, 1978), where $C(n) = \log(n)$, the natural logarithm of the number of observations n. When these criteria are used to select a model, the model with the lowest value of the information criterion is selected from the group of considered models.

The information criteria introduced in the previous paragraph don't take the model's intended use into account. Given the same dataset, and the same group of models to consider, they will always select the same model, irrespective of what the model will be used for in further steps. Claeskens and Hjort (2003), however, advocate a different approach. They assert that the model selection step should be dependent on the intended goal, and that it should allow to select that model which is the best for that particular goal, instead of a model which is overall reasonable, but not optimal for the chosen task. To this end, they define the *focus* of a model as a function $\mu(\beta)$ of the parameter β. This focus can be

any function which is piecewise continuous, for example, the predicted value of the response variable for a new observation x_0. Then, they define a *Focussed Information Criterion* as an unbiased estimate of the Mean Squared Error of the estimated focus $\mu(\hat{\beta})$, with $\hat{\beta}$, and use that to select an appropriate model. Once again, the model with the lowest value for the FIC is the model which is best for the given task.

In Chapter 2 we consider the application of the FIC for variable selection to the logistic regression model (1.5), where we have the specific goal of predicting the outcome y for new observations x_0. Recall that the Focused Information Criterion is defined as an unbiased estimate of the Mean Squared Error of the estimated focus. We extend this idea and define a more general FIC based on the L_p of the residuals, in other words, on

$$\frac{1}{n}\sum_{i=1}^{n}|r_i|^p$$

where p a positive integer. In addition, we also define a version of the Focused Information Criterion which is a direct estimate of the probability of misclassifying the new observation x_0, which is only possible because we work in a binary choice model setting. We illustrate the advantages of the original FIC, and our newly proposed variations of it, with a simulation experiment, and with a real data example. We find that using the FIC for variable selection results on average in models with a lower misclassification rate than models selected with established information criteria. This indicates that using different models for making predictions in different regions of the input space results in smaller/less errors than using just one model for the entire input space.

In Chapter 3 we continue to examine the properties of the FIC, though this time in the setting of stationary time series. Once again we concentrate on selecting models with a high predictive accuracy, in the sense of a low Mean Squared Error. Here the goal was to examine whether FIC shares the asymptotic efficiency property of the AIC, which has been proved in Shibata (1980), Bhansali (1996), Lee and Karagrigoriou (2001), and Ing and Wei (2005). Before this property could be examined, we first had to extend the theory of FIC so that

it allows the size of the largest model to increase as the number of observations, here the length of the considered time series, increases to infinity. We succeeded in making this extension by an adaptation of a theorem found in Portnoy (1985). Then, we illustrate, with both a simulation experiment and a real data example, that the FIC is a valid alternative for the established AIC and BIC for selecting the order of autoregression of a prediction time series model. Finally, we extend the presented ideas to non-predictive purposes, such as estimating the impulse response function of a time series, and we explore the use of the FIC in various extended time series models.

In Chapter 4 we once again examine the issue of variable selection in a binary choice model, but this time in the more recently developed Support Vector Machine setting. Despite the fact that SVMs work well in situations with a high number of explanatory variables, it has been demonstrated that even here a reduction of this number can increase the model's performance. We briefly examine the information criterion-based techniques which have already been developed, and find that they have the drawback of being computationally intensive. Therefor we propose two new information criteria which resemble the well-known AIC and BIC in the linear regression setting, and which have the advantage of being relatively fast to compute. We also demonstrate that one of these new criteria is a rough approximation of the recently developed Kernel Regularisation Information Criterion (Kobayashi and Komaki, 2006). Through a simulation study, we find that our new information criteria select models which give predictions at least as accurate (low misclassification rate) as the already developed criteria. We repeat this comparison on a few real benchmark datasets, and we arrive at the same conclusions.

For the last chapter, Chapter 5, we step away from the variable selection problem. Instead, we study the classification efficiency of the Support Vector Machine (SVM), and a few other Convex Risk Minimisation (CRM) methods. We first prove a few general properties about this class of binary classification techniques, and then we analyse each of them in more detail in the setting of two normally distributed populations with equal variances. In this setting, we know that the well known Fisher's Linear Discriminant rule is optimal, and we examine

how much of this efficiency is lost as a tradeoff for the additional flexibility given by the various CRM methods we examine. We calculate the classification efficiencies of these CRM techniques using influence functions (Hampel et al., 1986) as in Croux, Filzmoser and Joossens (2008), and we find that for reasonably balanced classes, the Convex Risk Minimisation techniques we studied are still quite efficient (efficiency $> 50\%$).

Finally, in the appendix we provide the proofs and the detailed analytical derivations of the results presented in Chapters 2 through 5.

Chapter 2

Variable Selection for Logistic Regression using a Prediction Focussed Information Criterion

This chapter is based on the following publication:
Claeskens, G., Croux, C. and Van Kerckhoven, J. (2006). Variable selection for logistic regression using a prediction focussed information criterion. *Biometrics*, **62**, 972–979.

Abstract

In biostatistical practice, it is common to use information criteria as a guide for model selection. We propose new versions of the Focussed Information Criterion (FIC) for variable selection in logistic regression. The FIC gives, depending on the quantity to be estimated, possibly different sets of selected variables. The standard version of the FIC measures the Mean Squared Error (MSE) of the estimator of the quantity of interest in the selected model. In this paper we propose more general versions of the FIC, allowing other risk measures such as one based on L_p-error. When prediction of an event is important, as is often the

case in medical applications, we construct an FIC using the error rate as a natural risk measure. The advantages of using an information criterion which depends on both the quantity of interest and the selected risk measure are illustrated by means of a simulation study and application to a study on diabetic retinopathy.

2.1 Introduction

Most clinical trials result in rich datasets with numerous variables of potential influence. Model selection methods are therefore becoming an essential tool for any data analyst. For an overview of model selection literature, see Burnham and Anderson (2002), George (2000), Spiegelhalter, Best, Carlin and van der Linde (2002) or Claeskens and Hjort (2003). In the Wisconsin Epidemiologic Study of Diabetic Retinopathy (WESDR; Klein et al, 1984) for example, there are eleven continuous covariates, amongst which are the duration of diabetes and the body mass index, and four binary explicative variables, such as the patient's gender, and the type of his/her area of residence. It is unlikely that all of these variables are important for all uses of the data. Outcome of interest in this study is the presence of retinopathy of any degree and we are in particular interested in the prediction of this event.

Traditional model selection methods such as AIC (Akaike, 1974) or BIC (Schwarz, 1978) select one subset of the covariates, no matter which use of the data will follow. The FIC, focussed information criterion (Claeskens and Hjort, 2003), on the other hand, is developed to select a set of variables which is best for a given focus. Hand and Vinciotti (2003) state that "in general, it is necessary to take the prospective use of the model into account when building it", and address explicitly the prediction problem. Given a patient's specific covariate information, the FIC selects a model that is best for, for example, predicting the presence of the disease of this particular patient. It might happen that one model is good for all patients, however, in the analysis of the WESDR we find different models for different patient groups. In particular, it turns out that the glycosylated hemoglobin level is more important, from a predictive point of view, for patients (both men and women) on a high-level insulin treatment than for

patients on a low-level insulin treatment.

The FIC in its original format interprets 'best' model in the sense of minimizing the mean squared error (MSE) of the estimator of the quantity of interest. A novel aspect of this paper is that we introduce focussed model selection based on different risk measures, and not only based on MSE. Especially in the context of prediction of an event, we propose and develop a new focussed information criterion based on the error rate as a risk measure.

In Section 2.3, we define this FIC based on the error rate, and give explicit formulae to compute it (see Section 2.3.1). In addition, we define a general FIC based on L_p-loss, and provide expressions for the most commonly used cases, in particular for the mean absolute error (MAE) for $p = 1$. For $p = 2$ we are back to the MSE results of Claeskens and Hjort (2003). Section 2.4 reports on a simulation study to assess the performance of the FIC, as compared to AIC and BIC. Section 2.5 applies the new model selection criteria to the WESDR data and some concluding remarks are made in Section 2.6.

2.2 Framework and notation

Assume that a set of data (x_i, y_i) is available, where x_i is a covariate vector of length $d + q$, containing the explicative variables which may be continuous or categorical, and y_i is a 0/1 response variable. The data are distributed according to the following model:

$$P(y_i = 1 \mid x_i) = F(x_i^t \beta) \quad \text{for} \quad 1 \leq i \leq n \tag{2.1}$$

where $F(\cdot)$ is the inverse logit function $F(u) = 1/\{1+\exp(-u)\}$, and $\beta = (\theta^t, \gamma^t)^t$ is the $(d + q)$-vector of parameters, where θ consists of the first d parameters, the ones that we certainly wish to be in the selected model, and γ holds the last q parameters, the ones that may potentially be included in the chosen model. While the expressions for the model selection criteria derived in this paper are obtained for logistic regression models, the ideas transfer immediately to other binary regression models.

Naturally, one can choose a complicated model that incorporates all the variables, even though usually only a few of them are significant. However, such a model is not guaranteed to give the best estimates of the quantity of interest. Adding more variables increases the total variability. Another issue with choosing a complex model is its lack of simplicity: medical researchers often prefer simple models, which are easier to interpret. The goal of this paper is to select a submodel of the logistic regression model (2.1), and to use that model to predict the value of the response variable for a "new" observation x_0.

The notation used in this paper is largely the same as in Claeskens and Hjort (2003), and the necessary quantities for defining the new FICs will be repeated here. In a local misspecification setting, we specify the true value of the parameter vector as $\beta_{\text{true}} = (\theta_{\text{true}}^t, \gamma_0^t + \delta^t/\sqrt{n})^t$, where n is the sample size and γ_0 is the value of γ for the "null model", i.e. the smallest model we consider, containing only the parameter θ. For the model described above, γ_0 is equal to zero. The *focus* parameter $\mu = \mu(\beta)$ is a function of the model parameters β. The linear predictor at a covariate value x_0 in the logistic model is an example of such a focus parameter, where $\mu(\beta) = x_0^t\beta$. The true value of the parameter of interest is then denoted by $\mu_{\text{true}} = \mu(\beta_{\text{true}})$.

For the model selection problem there are potentially 2^q estimators of $\mu(\beta)$ to consider, one for each subset S of $\{1, \ldots, q\}$. Other estimation methods, such as model averaging or shrinkage estimators, combine several of these submodel estimators. The model indexed by S contains the parameters θ and those γ_i for which $i \in S$. In practical applications, the user might rule out some of these subsets a priori. We denote γ_{0,S^c} the known vector of "null" values $\gamma_{0,i}$ for $i \in S^c$, the complement of S with respect to $\{1, \ldots, q\}$, and define $\hat{\mu}_S = \mu(\hat{\theta}_S, \hat{\gamma}_S, \gamma_{0,S^c})$ the maximum likelihood estimator of μ in the model indexed by S.

Let $J_{n,\text{full}}$ be the estimated $(d+q) \times (d+q)$ information matrix of the full model, and J_{full} the limiting information matrix. We assume that $J_{n,\text{full}}$ is of full rank, and denote its submatrices $J_{n,00}$, $J_{n,01}$, $J_{n,10}$ and $J_{n,11}$, corresponding to the dimensions of θ and γ respectively, and analogously for J_{full}. Since the model

used is a logistic regression model, straightforward calculations show that

$$J_{n,\text{full}} = -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\beta\partial\beta^t}\log f(Y_i \mid x_i,\beta) = \frac{1}{n}\sum_{i=1}^{n} p_i(1-p_i)x_i x_i^t, \tag{2.2}$$

with $f(\cdot)$ the binomial probability mass function, and $p_i = F(x_i^t\beta)$ the probability associated with observation i. For other choices of the inverse link function F, a different expression for $J_{n,\text{full}}$ results. In practice we insert for β in $J_{n,\text{full}}$ the full model estimator.

First define $K = J^{11} = (J_{11} - J_{10}J_{00}^{-1}J_{01})^{-1}$, the limiting variance of $\hat{\gamma}$ in the full model, and K_n its finite sample counterpart. Then we have

$$D_n = \hat{\delta}_{\text{full}} = \sqrt{n}(\hat{\gamma}_{\text{full}} - \gamma_0) \xrightarrow{d} D \sim \mathcal{N}_q(\delta, K), \tag{2.3}$$

where δ measures the distance between the null and true model (see Hjort and Claeskens (2003) for details and more discussion). The maximum likelihood estimator of μ in the model S has now the following limiting distribution (Hjort and Claeskens, 2003, Lemma 3.3)

$$\sqrt{n}(\hat{\mu}_S - \mu_{\text{true}}) \xrightarrow{d} \Lambda_S = \left(\frac{\partial\mu}{\partial\theta}\right)^t J_{00}^{-1}M + \omega^t(\delta - M_S K^{-1}D), \tag{2.4}$$

where $M \sim \mathcal{N}_d(0, J_{00})$ is statistically independent of D. Here we use the quantities $M_S = \pi_S^t(\pi_S K^{-1}\pi_S^t)^{-1}\pi_S$, the limiting variance of $(\hat{\gamma}_S, \gamma_{0,S^c})$, and $M_{n,S}$ its finite sample counterpart, and where π_S stands for the projection matrix of size $|S| \times q$, mapping any vector $\nu = (\nu_1, \ldots, \nu_q)^t$ to ν_S, the latter consisting of those ν_i for which $i \in S$. We also need the auxiliary vector $\omega = J_{10}J_{00}^{-1}\frac{\partial\mu}{\partial\theta} - \frac{\partial\mu}{\partial\gamma}$, where we evaluate the partial derivatives at the full model. For example, for the particular choice of parameter of interest $\mu(\beta) = x_0^t\beta$, these derivatives are $\frac{\partial\mu}{\partial\theta} = x_{0,0}$ and $\frac{\partial\mu}{\partial\gamma} = x_{0,1}$, where x_0 is partitioned according to the dimensions of θ and γ.

Some calculations yield that the limiting distribution Λ_S has mean and variance

$$\lambda_S = E[\Lambda_S] = \omega^t(I_q - M_S K^{-1})\delta, \tag{2.5}$$

$$\sigma_S^2 = \text{Var}(\Lambda_S) = \tau_0^2 + \omega^t M_S \omega, \tag{2.6}$$

with $\tau_0^2 = (\frac{\partial \mu}{\partial \theta})^t J_{00}^{-1} (\frac{\partial \mu}{\partial \theta})$ the variance of $\hat{\mu}_\emptyset$ in the null model, which is independent of S. Note that this distribution Λ_S is normal, with a non-zero mean due to the local misspecification setting.

The new FICs involve the mean and variance of the limiting distribution of Λ_S, given in (2.5) and (2.6). The expressions presented above are the theoretical values, assuming the limiting experiment is valid. In practice we need to estimate the information matrix of the full model $J_{n,\text{full}}$ and derive the needed components from this estimate. We estimate the vector δ by $\hat{\delta}_{\text{full}} = \sqrt{n}\hat{\gamma}_{\text{full}}$ as in (2.3). This leads, first, to maximum likelihood estimators of λ_S and σ_S^2, the mean and variance of the distribution Λ_S, in the model S and, second, to an estimator of the information criterion for the submodel S.

2.3 Prediction focussed information criteria

The traditional AIC and BIC information criteria are, as FIC, based on a likelihood approach. Where the FIC takes on different values, depending on a specified focus parameter, the AIC or BIC values do not depend on the purpose of the statistical analysis. In this section we show how the results of Claeskens and Hjort (2003) can be applied for obtaining focussed information criteria when prediction of a binary variable is of interest.

In Section 2.3.1 we derive the FIC taking as risk measure the error rate associated with the prediction of an event, tailored for logistic regression problems. In Section 2.3.2 we derive an expression for the FIC based on the L_p-error. We then verify this result with the FIC based on Mean Squared Error (MSE, $p = 2$) as obtained in Claeskens and Hjort (2003), and present the explicit expression for the FIC based on the Mean Absolute Error (MAE, $p = 1$). The expressions for the FIC based on L_p-risk hold in a general setting, but in the subsequent sections they are applied with the linear predictor of an observation, here the log-odds ratio, as the focus parameter: $\mu_{\text{true}} = x_0^t \beta_{\text{true}}$ and $\hat{\mu}_S = x_0^t \hat{\beta}_S$.

The selected model is then aimed at minimizing the L_p-loss when predicting the true value of the focus parameter.

For every considered submodel, indexed by S, the focussed information crite-

rion is computed and denoted by FIC_S. We select that subset S of $\{1, \ldots, q\}$ for which FIC_S is the smallest, this leads to the FIC-selected model which is indexed by the optimal S.

2.3.1 The FIC based on Error Rate

Our aim is to construct a selection criterion with the purpose of selecting the model that has the lowest probability of misclassifying a "new" observation x_0, assuming that it has been generated from the same model as the "training" data $\{(x_i, y_i) \mid 1 \leq i \leq n\}$. A natural choice for the risk function here, denoted $r_{\text{ER}}(S)$, is the probability of misclassifying the observation x_0. The abbreviation ER stands for Error Rate. Define y_0 the true response for an observation with covariates x_0 as a realization of the 0/1 random variable Y_0 with conditional probability $P(Y_0 = 1 \mid x_0) = F(x_0^t \beta_{\text{true}})$, and let $\hat{y}_{0,S}$ be the predicted response according to the model defined by S. Then,

$$r_{\text{ER}}(S) = P(Y_0 = 1 \text{ and } \hat{y}_{0,S} = 0 \mid x_0) + P(Y_0 = 0 \text{ and } \hat{y}_{0,S} = 1 \mid x_0).$$

Due to independence of Y_0 and $\hat{y}_{0,S}$, this expression reduces to

$$r_{\text{ER}}(S) = P(Y_0 = 1 \mid x_0)P(\hat{y}_{0,S} = 0 \mid x_0) + P(Y_0 = 0 \mid x_0)P(\hat{y}_{0,S} = 1 \mid x_0),$$

and hence, using the logistic regression model,

$$r_{\text{ER}}(S) = F(x_0^t \beta_{\text{true}})P(x_0^t \hat{\beta}_S < 0) + \{1 - F(x_0^t \beta_{\text{true}})\}P(x_0^t \hat{\beta}_S > 0).$$

This misclassification rate is only concerned with the sign of the estimated log-odds ratio, not with the actual value itself. We now apply the methodology of Claeskens and Hjort (2003), with $\mu_{\text{true}} = x_0^t \beta_{\text{true}}$ as focus parameter, and $\hat{\mu}_S = x_0^t \hat{\beta}_S$. We emphasize that our ultimate goal is prediction, rather than parameter estimation, and we only define a focus parameter for mathematical reasons, such that the results of Claeskens and Hjort (2003) can be applied.

We use Λ_S, the limiting distribution of $\sqrt{n}(\hat{\mu}_S - \mu_{\text{true}})$ as in (2.4), to approximate

$$P(x_0^t \hat{\beta}_S < 0) = P(\hat{\mu}_S < 0) = P\{\sqrt{n}(\hat{\mu}_S - \mu_{\text{true}}) < -\sqrt{n}\mu_{\text{true}}\}$$

by $\Phi\left\{-(\sqrt{n}\mu_{\text{true}} + \lambda_S)/\sigma_S\right\}$, with λ_S and σ_S^2 as in (2.5) and (2.6), and $\Phi(\cdot)$ the cumulative density function of the standard normal distribution. From this, the

following approximation is proposed for the risk function

$$r_{\mathrm{ER}}(S) \approx F(\mu_{\mathrm{true}})\Phi\left(\frac{-\sqrt{n}\mu_{\mathrm{true}} - \lambda_S}{\sigma_S}\right) + \{1 - F(\mu_{\mathrm{true}})\}\Phi\left(\frac{\sqrt{n}\mu_{\mathrm{true}} + \lambda_S}{\sigma_S}\right).$$

This risk measure serves as the basis for the *Focussed Information Criterion* based on *Error Rate*. Inserting the estimators, see Section 2, this leads to the FIC based on error rate

$$\mathrm{FIC}_{\mathrm{ER}}(S) = F(\hat{\mu}_{\mathrm{full}})\Phi\left(\frac{-\sqrt{n}\hat{\mu}_{\mathrm{full}} - \hat{\lambda}_S}{\hat{\sigma}_S}\right) + \{1 - F(\hat{\mu}_{\mathrm{full}})\}\Phi\left(\frac{\sqrt{n}\hat{\mu}_{\mathrm{full}} + \hat{\lambda}_S}{\hat{\sigma}_S}\right),$$

where we estimated μ_{true} by $\hat{\mu}_{\mathrm{full}} = \mu(\hat{\beta}_{\mathrm{full}})$. Note that this criterion depends on the value of the covariate vector x_0 of the observation to predict through the focus parameter μ, which is also present in the estimated values of λ_S and σ_S, see (2.5) and (2.6).

2.3.2 The FIC based on L_p-error

Based on the limiting distribution of $\sqrt{n}(\hat{\mu}_S - \mu_{\mathrm{true}})$ in equation (2.4), we derive the expressions for the L_p-error of $\hat{\mu}_S$, and this for any subset S of $\{1, \ldots, q\}$ and for any positive $p \geq 1$. The L_p-risk measure is defined as the p^{th} order absolute moment of the limiting distribution Λ_S, $r_p(S) = E(|\Lambda_S|^p)$. Note that we work with the absolute moments and not the centered ones because we want a measure of the deviations of $\hat{\mu}_S$ to μ, and the bias involved should not be eliminated by centering. For integer values of p it is possible to derive an explicit expression for $r_p(S)$. The general expressions, and details on their derivation, can be found in Appendix A.1. Note again the dependence of $r_p(S)$ on the focus parameter: different choices of μ will lead to different formulae for the focussed criterion, and as a consequence, may lead to different selected models.

We now give details on two special cases of the FIC based on L_p-error. The first case is FIC_2 based on the L_2-error, better known as the mean squared error and henceforth denoted as $\mathrm{FIC}_{\mathrm{MSE}}$. This model selection criterion has been extensively discussed in Claeskens and Hjort (2003). For $p = 2$, $r_2(S) = \lambda_S^2 + \sigma_S^2$. Applying equations (2.5) and (2.6), this can be written as

$$r_2(S) = \omega^t(I_q - M_{n,S}K_n^{-1})\delta\delta^t(I_q - K_n^{-1}M_{n,S})\omega + \tau_0^2 + \omega^t M_{n,S}\omega, \qquad (2.7)$$

which is, up to a constant term, equal to the limit FIC as defined in Claeskens and Hjort (2003). Note that an asymptotically unbiased estimate of $\delta\delta^t$ in (2.7) is given by $\hat{\delta}\hat{\delta}^t - K_n$. Inserting unbiased estimators leads to

$$\mathrm{FIC}_{\mathrm{MSE}}(S) = \hat{\omega}^t(I_q - M_{n,S}K_n^{-1})\hat{\delta}\hat{\delta}^t(I_q - K_n^{-1}M_{n,S})\hat{\omega} + 2\hat{\omega}^t M_{n,S}\hat{\omega}.$$

The other special case that we study is $p = 1$, which leads to a "new" criterion minimizing the mean absolute error, MAE. Here it can be verified that

$$r_1(S) = 2\lambda_S\left\{\Phi(\frac{\lambda_S}{\sigma_S}) - \frac{1}{2}\right\} + 2\sigma_S\phi(\frac{\lambda_S}{\sigma_S}).$$

Then we define the Focussed Information Criterion based on MAE as the following estimator of $r_1(S)$

$$\mathrm{FIC}_{\mathrm{MAE}}(S) = 2\hat{\lambda}_S\left\{\Phi(\frac{\hat{\lambda}_S}{\hat{\sigma}_S}) - \frac{1}{2}\right\} + 2\hat{\sigma}_S\phi(\frac{\hat{\lambda}_S}{\hat{\sigma}_S}),$$

where $\phi(\cdot)$ is the density function of the standard normal.

2.4 Simulation study

In this section, a simulation study is presented to examine how well the proposed Focussed selection criteria perform with respect to two better known criteria, the Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC). In Section 2.4.1, the particulars of the simulation sampling scheme are detailed. In Section 2.4.2 we additionally address the issue of model averaging. The results of the simulation are presented in Section 2.4.3.

2.4.1 Simulation settings

For the simulation study, $n_{\mathrm{test}} = 500$ observations $x_{0,i}$ are independently generated from a normal $\mathcal{N}_q(0, \frac{1}{4}I_q)$ distribution, with I_q the $q \times q$ identity matrix. These observations constitute the test sample and remain the same throughout the entire simulation. Then, for each of the $M = 1000$ simulations in the experiment, a training sample of size n_{train} observations (x_i, y_i) is generated, according

to the model

$$P(y_i = 1 \mid x_i) = F(\theta + x_i^t \gamma),$$

where $\theta = 0$, $\gamma = (1, -1, 1, -1, 0, \ldots, 0)^t$ such that only 4 out of the q covariates are pertinent. Again, $x_i \sim \mathcal{N}_q(0, \frac{1}{4} I_q)$, where the factor $\frac{1}{4}$ is present so that the generated linear predictors $x_i^t \beta$ are distributed according to a standard normal distribution. For each simulation run, we minimize the information criterion under investigation, and force the intercept term to be in every model. Within each simulation run, AIC and BIC select one single best model, while for each one of the n_{test} observations in the test sample, possibly different models according to FIC_{MSE}, FIC_{MAE} and FIC_{ER} are selected. The forward search method as described in Section 2.4.2 has been used, and in each of those selected models we use the estimator $\hat{\mu}_{0,i} = \hat{\theta} + x_{0,i}^t \hat{\gamma}$. Its sign determines the predicted value of the corresponding binary $y_{0,i}$ values. We did experiments with $n_{\text{train}} = 50$ and 200, and $q = 5$ and 9.

For each separate observation $x_{0,i}$ in the test sample, with $1 \leq i \leq n_{\text{test}}$, we measure the performance of the model selection criteria via (a) the mean squared error of $\hat{\mu}_{0,i}$, (b) its mean average deviation, and (c) the error rate. The MSE is given by

$$\text{MSE}(\hat{\mu}_{0,i}) = \frac{1}{M} \sum_{j=1}^{M} (\hat{\mu}_{0,i}^{(j)} - \mu_{0,i,\text{true}})^2,$$

with $\hat{\mu}_{0,i}^{(j)}$ the estimated value for validation observation $x_{0,i}$ in simulation run j, and $\mu_{0,i,\text{true}}$ the corresponding true value. Similarly, the MAE is computed as

$$\text{MAE}(\hat{\mu}_{0,i}) = \frac{1}{M} \sum_{j=1}^{M} |\hat{\mu}_{0,i}^{(j)} - \mu_{0,i,\text{true}}|.$$

The MAE performance measure is sometimes preferred since it is, compared to MSE, less influenced by those simulation runs yielding large deviations from the true values. Finally, the error rate is simulated as

$$\text{ER}_i = \frac{1}{M} \sum_{j=1}^{M} I(\hat{\mu}_{0,i}^{(j)} \mu_{0,i,\text{true}} < 0)$$

where $I(\cdot)$ is the indicator function. If the estimated and the true linear predictor have the same sign, they give a zero contribution to the sum in the above ER_i. Otherwise, they contribute to the error rate.

2.4.2 Further particulars

A search across all possible models is only feasible for q relatively small, because the number of possible models to search through increases exponentially with q. A forward selection approach is an alternative to an exhaustive search, possibly leading to a different selected model. Starting from the null model, this iterative procedure adds one variable at a time. Specifically, it adds that variable which yields the lowest value for the information criterion when added to the currently "best" model. This process is repeated until $q+1$ nested models are obtained, ranging from the null model to the full model and indexed by $S_0, S_1, \ldots, S_q$. From these models, we select the model that yields the lowest value for the information criterion. Alternatively, we can apply a backward elimination procedure, starting with the full model, and eliminating in each step the variable which gives the largest reduction (or smallest increase) to the value of the information criterion. This will also lead to $q+1$ nested models as described above, from which we choose the model with the lowest value of the information criterion.

Model averaging can be applied as an alternative to selecting a single model (see also Hjort & Claeskens (2003)). In this case we construct a weighted average of the estimators in the different models. For each of the nested models obtained during the forward variable selection procedure, we compute the weight as

$$w_j = \frac{\exp\{-\frac{1}{2}\text{xIC}(S_j)\}}{\sum_{k=0}^{q}\exp\{-\frac{1}{2}\text{xIC}(S_k)\}}$$

where $\text{xIC}(S_k)$ is the value of the Information Criterion (AIC, BIC, FIC, ...) at the model S_k with k included variables, for $k = 0, \ldots, q$. For each of the submodels S_j a prediction of $\mu_0 = x_0^t\beta$ for an observation to be classified, is obtained, and these predicted values $\hat{\mu}_{0,S_j}$ then generate the "model-averaged" prediction $\hat{\mu}_0 = \sum_{j=0}^{q} w_j\hat{\mu}_{0,S_j}$. The advantage of a model averaged estimator is that it might have reduced variability. This will be illustrated in the simulation

experiments, where results for the "model-averaged" procedure are reported as well. In the classification literature it is a common strategy to combine several classifiers, see, e.g., Kuncheva (2004) for an overview. Of course, averaging over all possible subsets of the full model, or over any other sequence of models is possible.

All computations are performed using the publicly available software package R. In our software we define $\text{AIC}_S = -2\log L(\hat{\beta}_S) + 2(p + |S|)$, and similarly $\text{BIC}_S = -2\log L(\hat{\beta}_S) + \log(n_{\text{train}})(p + |S|)$, with $L(\hat{\beta}_S)$ the likelihood of the estimated model indexed by S, and $|S|$ the number of elements in the subset S, such that lower values indicate better models.

2.4.3 Simulation results

This simulation results in $n_{\text{test}} = 500$ distinct values of the MSE, MAE and Error Rate, one for each observation in the test sample, for prediction based on a submodel selected by AIC, BIC, FIC_{MSE}, FIC_{MAE}, and FIC_{ER}. These values are also computed for the model-averaged predictions, discussed in Section 2.4.2. For the case $n_{\text{train}} = 50$ and $q = 5$, Table 2.1 presents the averages, after applying the log-transform to MSE and MAE, of the performance measures over the $n_{\text{test}} = 500$ values, together with their standard error (SE). The log-transformation is applied to the MSE and the MAE, to make their distributions more symmetric. The boxplots in Figures 2.1 and 2.2 provide a graphical representation of these 500 values.

First of all, we see from Table 2.1 that model averaging significantly improves the performance for the MSE and MAE. In terms of Error Rate, model averaging does not seem to give much improvement, but neither a worsening of the results obtained with single model selection. We see that FIC_{ER} gives the best results for the Error Rate, FIC_{ER} selects, compared to the other selection criteria, the models which yield the lowest error rates. This should not be too surprising, since the risk measure associated with FIC_{ER} is the error rate (to be more precise, the error rate of the limiting experiment), and FIC_{ER} selects the model having the smallest value of an approximation of this risk measure. It can be verified that

	log(MSE)		log(MAE)		Error Rate ($\times 10^{-2}$)	
Criterion	Average	SE	Average	SE	Average	SE
AIC	0.141	0.025	-0.182	0.013	26.62	0.60
BIC	0.153	0.027	-0.152	0.014	33.89	0.48
FIC_{MSE}	-0.026	0.024	-0.298	0.013	24.65	0.64
FIC_{MAE}	0.085	0.024	-0.238	0.012	22.87	0.65
FIC_{ER}	0.507	0.024	0.034	0.013	20.75	0.65
aAIC	0.045	0.025	-0.238	0.013	25.45	0.62
aBIC	0.025	0.026	-0.226	0.014	31.14	0.55
$a\text{FIC}_{\text{MSE}}$	-0.402	0.021	-0.438	0.011	24.23	0.64
$a\text{FIC}_{\text{MAE}}$	-0.454	0.021	-0.467	0.011	22.34	0.64
$a\text{FIC}_{\text{ER}}$	-0.220	0.023	-0.341	0.013	20.91	0.64
full model	0.065	0.024	-0.253	0.012	20.75	0.65

Table 2.1: Average values, together with their standard errors (SE), of the log(MSE), log(MAE) and Error Rates over the 500 observations to predict in the test sample for the sampling scheme with $n_{\text{train}} = 50$ and $q = 5$. The MSE, MAE, and Error rates have been simulated for estimators of a model selected by the criteria AIC, FIC_{MSE}, FIC_{MAE}, and FIC_{ER}, as well as for the model averaged versions of the estimators (indicated by the prefix "a").

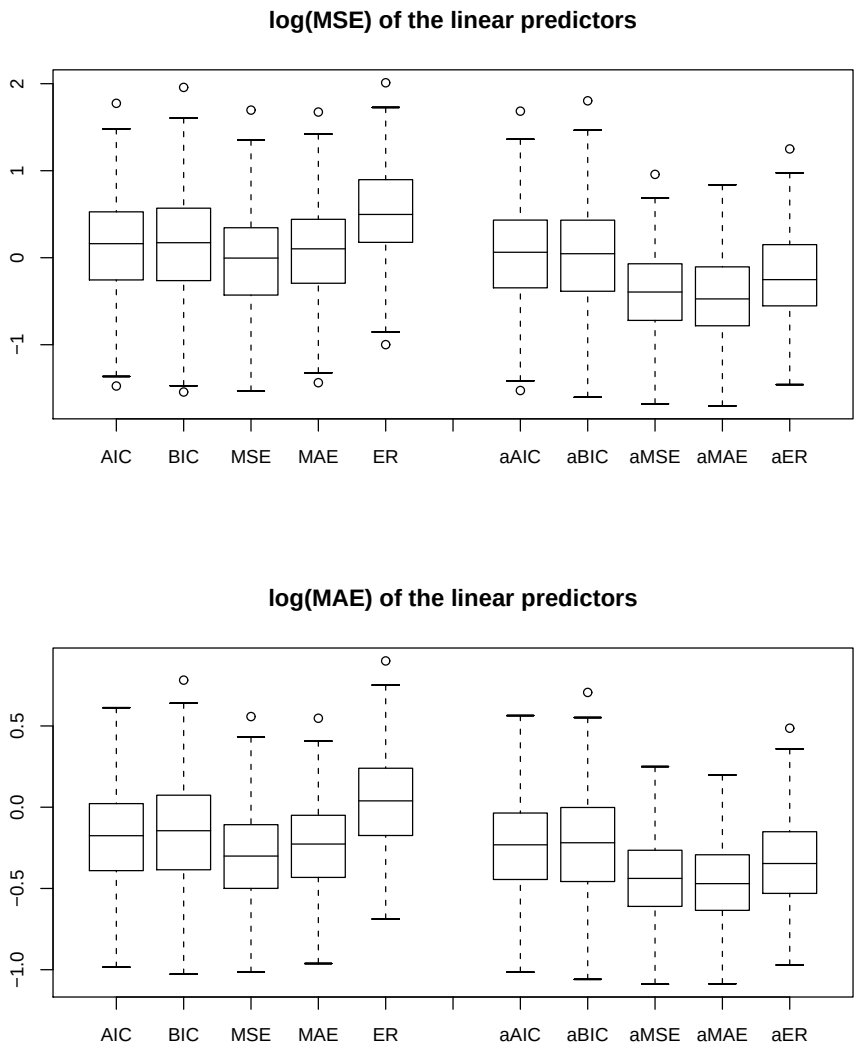

Figure 2.1: Boxplots of the log(MSE) and log(MAE) of the 500 observations to predict in the test sample for the sampling scheme with $n_{\text{train}} = 50$ and $q = 5$. The MSE and MAE have been simulated for estimators of a model selected by the criteria AIC, BIC, FIC_{MSE}, FIC_{MAE}, or FIC_{ER}, as well as for the model averaged versions of the estimators (indicated by the prefix "a").

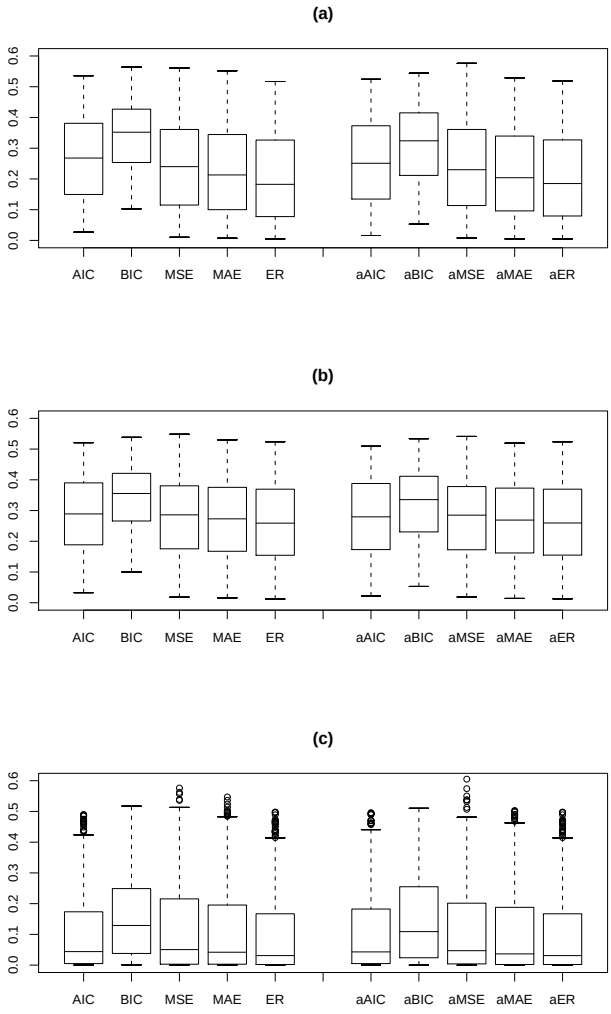

Figure 2.2: Boxplots of the Error Rates of the 500 observations to predict in the test sample. These Error Rates have been simulated for estimators of a model selected by the criteria AIC, BIC, FIC_{MSE}, FIC_{MAE}, or FIC_{ER}, as well as for the model averaged versions of the estimators (indicated by the prefix "a"). In the top panel (a) $n_{\text{train}} = 50$, and $q = 5$ variables, in (b) $n_{\text{train}} = 50$, and $q = 9$ variables, and in panel (c) $n_{\text{train}} = 200$, and $q = 5$ variables.

the average Error Rate of the FIC_{ER} is indeed significantly smaller than the other average error rates reported in Table 2.1, both for single model predictions and for averaged-model predictions. The average error rates are computed over n_{test} outcomes, and differences among them have been tested for by performing multiple paired comparisons tests with Tukey's Honest Significant Difference method (e.g., Neter et al., 1996, page 725-732) and resulted in P-values < 0.01. Also, comparing with the results from the full model, given in the bottom line in Table 2.1, we see that FIC_{MSE} outperforms the full model in terms of MSE and MAE, and that FIC_{ER} does as good as the full model in terms of Error Rate. The models selected by FIC_{ER} however, generally have a small number of selected variables, and hence are much easier to interpret than the model which includes all variables.

The plots in Figure 2.1 show that FIC_{MSE} and FIC_{MAE} outperform the selection procedure based on AIC and BIC when using MSE and MAE as performance criterion. Again, one can show that these differences in average performance are also highly significant, and become after model-averaging even more pronounced. This is as one should expect, since variable selection using FIC_{MSE} and FIC_{MAE} is aimed at choosing the "best" model as measured by the risks MSE and MAE. While FIC_{ER} gives the best results for the Error Rate performance criterion, it performs comparatively much worse for MSE and MAE. But this should not be of much concern, since if the researcher thinks that another risk measure than Error Rate is more appropriate for his/her prediction problem, he/she should use a variable selection method focussed on that particular risk function.

Comparing FIC_{MSE} and FIC_{MAE} is more difficult. When selecting a single model, the MAE for estimates based on FIC_{MAE} is on average slightly worse than for FIC_{MSE}, although the difference is only minor. Note that at the finite-sample level there is no guarantee that the model selected using the FIC_{MAE} indeed yields the smallest Mean Absolute Errors. Moreover, the FIC is only estimating the limiting risk measures, and uncertainty from estimating population quantities needs to be taken into account. Most important, however, is that in this simulation setting, both FIC_{MSE} and FIC_{MAE} do better than AIC and BIC, both for model selection and model averaging.

Our simulations also indicated that increasing the number of variables q to 9, or increasing the training sample size to 200 does not change the above conclusions. Of course, for $n_{\text{train}} = 200$ all MSE/MAE will be lower than for a training sample size of 50. In Figure 2.2, boxplot representations of the n_{test} simulated error rates are given for the cases (i) $n_{\text{train}} = 50$ and $q = 5$ (ii) $n_{\text{train}} = 50$ and $q = 9$ and (iii) $n_{\text{train}} = 200$ and $q = 5$. Again we observe that FIC_{ER} performs the best on this criterion, especially for small training sample sizes ($n_{\text{train}} = 50$), and this remains true if we apply model averaging. We also observe that for the larger training sample sizes ($n_{\text{train}} = 200$), the performances of the different model selection methods are closer together. This is again as expected, since if n_{train} gets larger, the variance of the parameter estimators decreases.

2.5 Analysis of WESDR data

In this section we perform model selection for the 1998 data of the Wisconsin Epidemiologic Study of Diabetic Retinopathy (WESDR), with the methods described in Section 2.3. The data consists of 691 records of subjects with younger-onset diabetes (the incomplete observations were removed before the analysis). The response variable 'y' is a 0/1 variable where 1 indicates the presence of retinopathy of any degree. The 11 continuous covariates are 'rere' and 'lere', the refractive error in diopters for respectively the right and the left eye; 'reip' and 'leip', the internal eye pressure in mmHg for respectively the right and the left eye; 'adia', the age in years at which diabetes was diagnosed; 'ddia', the duration of diabetes in years; 'gly', the percentage of glycosylated hemoglobin, 'sysp' and 'diap', the systolic and diastolic blood pressure in mmHg; 'bmi', the Body Mass Index, and 'pulse', the pulse rate in beats per 30 seconds. The 4 binary 0/1 covariates are 'sex', with 1 indicating male; 'uri', with 1 indicating the presence of urine protein; 'ins', with 1 indicating more than 1 dose of insulin taken per day, and 'urb', with 1 indicating that the subject lives in an urban county.

When we fit a model including all the variables, we find that the following are significant at the 5% level: 'ddia', 'gly', 'urb' (in decreasing order of significance). Some pairs of variables are strongly correlated, for example 'lere' and 'rere' (with

$r = 0.869$), and 'reip' and 'leip' ($r = 0.872$). These four variables are also the ones with the largest Variance Inflation Factor (above the critical value 3), as computed by the `R` software package, following Davis, Hyde, Bangdiwala, and Nelson (1986), pp. 140–147. We refer to Klein et al. (1984) for further discussion of this data set. Given the high number of variables and the correlations among them, we want to select a subset of variables, most pertinent for predicting the response variable for a new patient.

We examine the predictive power of the models selected by the different selection criteria AIC, BIC, FIC_{MSE}, FIC_{MAE}, FIC_{ER}, as well as the model-averaged version by assessing their error rates. Since the total number of all possible submodels amounts to 2^{15}, we carried out the model selection using a forward search procedure, as discussed in Section 2.4.2, to speed up the computation time. Also note that, since we work with real data for which the true value of the linear predictors is not available, the MSE and MAE performance criteria cannot be computed. The error rate is estimated by means of a cross-validation experiment: for each patient in the dataset, we select and estimate a model based on all the other patients in the dataset and then make a prediction for the presence of retinopathy of the left-out observation. Then, we compare the predictions with the real values of 'y', the presence of retinopathy of any degree. We count the percentage of wrong predictions, which yields an estimate of the error rate. The results are summarized in Table 2.2.

Method	AIC	BIC	FIC_{MSE}	FIC_{MAE}	FIC_{ER}
Error rate (no model averaging)	0.198	0.184	0.174	0.174	0.177
Error rate (after model averaging)	0.194	0.188	0.171	0.174	0.174

Table 2.2: Error rates for the WESDR data, obtained via cross-validation. The models are selected using AIC, BIC FIC_{MSE}, FIC_{MAE} FIC_{ER} and also results for the model-averaged estimates are reported.

We observe from Table 2.2 that the models selected by the focussed information criteria and the model-averaged estimates based on FIC, all yield a lower error rate than their AIC and BIC counterparts. The McNemar test (e.g. Kuncheva 2004, page 13-15) reveals that in particular the difference with the AIC-selected model is strongly significant (P-values < 0.025). On the other hand, the difference between the error rates for the models selected by the different FICs is not statistically significant. These results illustrate the advantage of selecting a possibly different set of predictor variables for every observation to predict. Indeed, there is a priori no reason why a unique selected model would be best for all future predictions to be made. If the "right" model would be within the class of allowed models, then this is presumably the best model to use for prediction. However, we do not believe that the "right" model does exist, only that some models are better than others, depending on the purpose of the analysis.

To illustrate that the model selected by the FIC might depend on the observation, we performed a second analysis. We divided the patients into four groups, according to their gender and the number of doses of insulin taken each day, as shown below.

Group	characteristics
A	females taking none or a single insulin dose each day
B	females taking multiple insulin doses each day
C	males taking none or a single insulin dose each day
D	males taking multiple insulin doses each day

The groups have roughly an equal number of observations. We record for each group the percentage of times that each variable enters the model when predicting an observation belonging to that group. Table 2.3 shows the selection frequencies for the four most often selected variables in every group, for FIC_{MSE} and FIC_{ER}.

The FIC methods select the variable 'ddia' most often, and in particular the error rate based FIC has a strong preference for this variable. A logistic regression model containing only an intercept and this variable 'ddia' performs very well, with a cross-validated error rate of 0.189. In fact, the model selected using FIC_{ER} ends up with this simple model in 46.3% of the cases. But, as follows from Table

	Group	Variable 1		Variable 2		Variable 3		Variable 4	
FIC_{MSE}	A	ddia	86.2%	gly	53.8%	pulse	42.6%	reip	39.0%
	B	ddia	81.8%	gly	50.0%	pulse	33.8%	urb	32.4%
	C	ddia	78.5%	gly	51.3%	pulse	34.4%	reip	33.8%
	D	ddia	77.8%	gly	54.9%	reip	39.2%	pulse	37.9%
FIC_{ER}	A	ddia	92.3%	gly	28.2%	reip	17.4%	uri	16.9%
	B	ddia	90.5%	gly	45.3%	uri	33.8%	diap	25.0%
	C	ddia	89.2%	gly	36.4%	uri	31.8%	bmi	24.6%
	D	ddia	90.8%	gly	41.8%	uri	32.0%	pulse	28.8%
AIC		ddia	yes	gly	yes	bmi	yes	pulse	yes
BIC		ddia	yes	gly	yes	bmi	no	pulse	no

Table 2.3: Model selection methods FIC_{MSE} and FIC_{ER} are applied to each subject within a group of the WESDR data. The table shows the selection percentages of the four most frequently selected variables per group. For completeness, the last 2 rows show the first four variables considered for inclusion by AIC and BIC, and whether they have been selected ("yes") or not ("no").

2.2, the FIC_{ER} approach reaches even a lower error rate by deviating from this simple model for an important part of the observations to classify. A possible strategy for a more refined analysis is to add the variable 'ddia' in the list of fixed variables which are included in every selected model, together with the intercept.

The second most selected variable is 'gly', the percentage of glycosylated hemoglobin, which is selected about half of the time by the FIC based on MSE, and with a lower frequency by the FIC based on error rate. Fitting a logistic regression model containing only the intercept, 'ddia' and 'gly', we find a cross-validated error rate of 0.184, still above the error rates found with the focussed information criteria. (Note that adding the third most significant variable, 'urb', does not further improve the error rate). In Table 2.3, the variables being selected first in the forward procedure by AIC and BIC are also reported. We see that BIC only selects 'ddia' and 'gly', while the model finally selected by the AIC criterion contains 7 variables.

Variable selection based on FIC_{ER} includes the variable 'gly' much more often for groups B and D than for groups A and C (see Table 2.3). Hence, there is some indication that the glycosylated hemoglobin level is, from a predictive point of view, less important for patients taking none or only a single dose of insulin each day (groups A and C) than for patients taking multiple doses of insulin each day (groups B and D). If a full model approach is opted for, it might be advisable to include an interaction term between the two variables 'gly' and 'ins'.

2.6 Conclusions

In this paper, we extended the focused information criterion, as developed by Claeskens and Hjort (2003). It is originally constructed to select a submodel minimizing the mean squared error of the estimator of the focus point. The idea put forward in this paper is that MSE is not the only risk measure that one can consider. We expand the construction and application to minimize the more general L_p-norm, of which MSE ($p = 2$) and mean absolute deviation ($p = 1$) are special cases. Another contribution of this paper is the proposal of a Focussed Information Criterion using the error rate as risk measure. This is of specific use

in binary regression problems, where the goal is to select models which yield the lowest error rate.

To show the usefulness of these information criteria, we presented both a simulation study and an analysis of the WESDR dataset. In these analyses, we observed that the focussed information criteria select models which perform better with respect to their specific risk measure (that is, lower MSE for the FIC based on MSE, and lower error rate for the FIC based on error rate), than the Akaike information criterion. In the WESDR data analysis, it was illustrated how different models are selected for different patients. By allowing the selected model to vary with the observation to predict, a gain in predictive performance is expected.

The variable selection problem becomes even more pertinent when a large number of variables relative to sample size is available. In this setting, the non-existence of the classical logistic regression estimator may cause problems. It is a topic of our current research to apply model selection methods to such data sets.

Chapter 3

Prediction Focussed Model Selection for Autoregressive Models

This chapter is based on the following publication:
Claeskens, G., Croux, C. and Van Kerckhoven, J. (2007). Prediction focussed model selection for autoregressive models. *Australian and New Zealand Journal of Statistics*, **49**, 359–379.

Abstract

In order to make predictions of future values of a time series, one needs to specify a forecasting model. A popular choice is an autoregressive time series model, where the order of the model is chosen by an information criterion. We propose an extension of the Focussed Information Criterion (FIC) for model-order selection with focus on a high predictive accuracy (i.e. the mean squared forecast error is low). We obtain theoretical results and illustrate via a simulation study and some real data examples that the FIC is a valid alternative to the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) for selection of

a prediction model. We also illustrate the possibility of using FIC for purposes other than forecasting, and explore its use in an extended model.

3.1 Introduction

In many fields of applied research (e.g. economics, demographics), a variable is observed over time, and the researcher wishes to model the time-structure of the data and predict future values of the variable. This modelling consists of two important parts: first, the general trend over time is modelled and seasonal effects are identified, and then the dynamic structure of the resulting stationary series is investigated. In this paper we are mainly concerned with the latter. A popular choice is the autoregressive model

$$Z_t = \phi_1(p)Z_{t-1} + \cdots + \phi_p(p)Z_{t-p} + \varepsilon_t(p), \tag{3.1}$$

which predicts the stationary variables Z_t by its lagged variables. Model (3.1) is an autoregressive model of order p, abbreviated as an AR(p)-model. The variables Z_t have been centred by their average, and the $\varepsilon_t(p)$ are zero mean, white noise innovation terms. Modelling the time series can serve many purposes, but usually the goal is to make accurate predictions of the series in the unobserved future.

We focus on making forecasts of the series h steps beyond the last observation. Generally, the accuracy of these forecasts depends on the autoregressive order p of the model used, in other words on how far in the past we look in order to model the series. If we restrict ourselves to only the recent past, p small, then we might fail to capture more long-term influences. Conversely, if we include the far past, p large, then the accuracy of the predictions will suffer because of the chosen model's complexity. Hence, a balance between completeness and simplicity must be chosen, and a commonly used method of selecting an appropriate AR-order is by computing the value of an information criterion for each candidate model, and selecting the model with the best value of the criterion.

In this paper, we propose an adapted version of the Focussed Information Criterion (FIC) as defined in Claeskens and Hjort (2003). The main novel aspects are the application to time series and that we allow the maximal order of

the autoregressive model to increase slowly to infinity as the length of the series increases. We also provide a bound on the rate of this increase by adapting a theorem in Portnoy (1985) to the time series setting. This result is needed because, originally, the theory behind FIC was developed for the case where the maximal number of variables in the model, or in this case the maximal considered autoregressive order, remains constant. We develop these ideas in the setting of two independent realizations of the data generating process, hereby following Shibata (1980), Bhansali (1996), and Lee and Karagrigoriou (2001). This framework is described in Section 3.2, where we also discuss the more realistic case of only one realization of the data generating process. Section 3.3 contains the derivation of the FIC.

In Section 3.4 we report the results of a simulation study. We compare the efficiency in mean squared error sense of the models selected by FIC with the efficiency of two well-known criteria: Akaike Information Criterion (AIC; Akaike, 1974) and Bayesian Information Criterion (BIC; Schwarz, 1978), also sometimes called Schwarz Information Criterion (SIC). First, the single-series setting is discussed, where AIC has recently been proven to be an asymptotically efficient criterion (Ing and Wei, 2005). We also performed a simulation study in the two-series setting (Shibata, 1980; Bhansali, 1996; Lee and Karagrigoriou, 2001), and compared it to the single-series setting. We expect FIC to perform well in this setting since FIC is constructed to minimise the estimated Mean Squared prediction error.

To illustrate the practical use of FIC, we compare in Section 3.5 the performance of the aforementioned criteria on two real data examples. In Section 3.6, we provide some extensions to the ideas presented in this paper, such as the application of the FIC to simultaneously select a subset of regression variables and the autoregressive order of the error terms, as in Shi and Tsai (2004). Finally, we summarise and make some concluding remarks in Section 3.7.

3.2 Model setting

In this section we state the model setting, and define h-step ahead predictions of a time series. The true time series is a realisation of an AR(∞)-process, and we approximate this by a finite order autoregressive model. We first assume that we have a univariate time series $\{y_t\}$ available, where $t = 1, \ldots, T$, and that we want to make a prediction of this series at time-horizon h. We denote this prediction $\hat{y}_{T+h}$. We also assume that we have a second series $\{x_t\}$ available of the same length T. This is the setting as used in Shibata (1980), Bhansali (1996), and Brockwell and Davis (1995, page 301), where statistical properties of model selection methods in time series are discussed. The two series are assumed to be independent realisations of the same length T of a stochastic process $\{Z_t\}$, with the following dependency structure:

$$Z_t = \varepsilon_t + \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \cdots . \tag{3.2}$$

We assume that the innovation terms ε_t are independent and identically normally distributed, with mean 0 and variance σ^2. We also assume that the autoregression coefficients ϕ_i are absolutely summable (that is $\sum_i |\phi_i| < \infty$), and that the associated power series

$$\Phi(z) = 1 - \phi_1 z - \phi_2 z^2 - \cdots$$

converges and is different from zero for $|z| \leq 1$. Our goal is to select the best approximating autoregressive model of order p, with $0 \leq p \leq p_T$, using the series $\{x_t\}$. Here we allow the maximal considered AR-order, denoted by p_T, to depend on T. This is done because one typically fits a time series model of a higher order if the length of the series is increased. Next, we use this selected model to make a h-step ahead forecast for the series $\{y_t\}$.

Although the two-series setting may seem artificial, and only of use for mathematical convenience, there are some cases where it can be considered to hold. Suppose that one is in a process control situation, where the performance of a machine is measured at regular time intervals. When the process is under control, a benchmark sample $\{x_t\}$ for the machine's performance can be taken, and the

researcher can fit a model to these data. At a later stage, another set of readings $\{y_t\}$ is taken. Based on these readings, and using the model found with the benchmark data, the next value is predicted. This prediction is then compared to the realised value and a large deviation could signify a problem with the machine.

In most practical situations, however, the user has only one time series $\{x_t\}$ available. In such a case, we make a h-step ahead prediction of the series $\{x_t\}$ itself. Our results are valid for both situations: one series and two series. For notational simplicity, we continue to work in the two-series setting. The results for the single-series setting are obtained by setting $\{y_t\}$ equal to $\{x_t\}$.

There are two methods to make the h-step ahead forecast $\hat{y}_{T+h}$. The first method is the direct method, which assumes that we estimate different models

$$Z_t = \phi_1(p,h)Z_{t-h} + \cdots + \phi_p(p,h)Z_{t+1-h-p} + \varepsilon_t(p,h) \tag{3.3}$$

for each horizon h. The $\varepsilon_t(p,h)$ are assumed to have zero mean and variance $\sigma^2(p,h)$. We forecast the series $\{y_t\}$ at horizon h by $\hat{y}_{T+h} = \hat{\phi}_1(p,h)y_T + \cdots + \hat{\phi}_p(p,h)y_{T+1-p}$. Here, the parameters $\phi_i(p,h)$ are estimated using ordinary least squares (OLS). This would make little difference as opposed to using the maximum likelihood (ML) estimator, especially for large T, while the ML-estimator would complicate the computations. The second method is the plug-in method. This is the more common approach and follows immediately from the estimates of model (3.2). Here, we compute recursively

$$\hat{y}_{T+h}(p) = \hat{\phi}_1(p)\hat{y}_{T+h-1}(p) + \cdots + \hat{\phi}_p(p)\hat{y}_{T+h-p}(p) \tag{3.4}$$

with $\hat{y}_t(p) = y_t$ for $t \leq T$. Once again, the parameter estimates $\hat{\phi}_i(p)$ are obtained using OLS. Observe that both methods are identical for $h = 1$. In the main part of this paper, we make predictions using the direct method; however, see Section 3.6.1 for the plug-in method. The main advantage of using the direct method and not the plug-in method was shown in Bhansali (1996). He showed that the lower bound on the Mean Squared Error (MSE) of predictions obtained via the direct method method is lower than that of the plug-in method. Also, he showed that, for the direct method, this lower bound can be achieved, which is not the case for the plug-in method.

3.3 The focussed information criterion

In this section we propose an extended version of the FIC as defined in Claeskens and Hjort (2003). The idea of the FIC is that an information criterion should take into account the purpose of the statistical analysis, by trying to estimate the MSE of the estimator of a focus parameter. For example, Claeskens et al. (2006) used the predicted value in a logistic regression model as a focus parameter. In the setting of this paper, the focus parameter is the h-step ahead prediction of a time series. In this extension, we allow the number of variables to increase towards infinity with the sample size. In time series analysis we select an AR(p)-model that fits the available data best, with $0 \leq p \leq p_T$. Recall that p_T is the maximal autoregressive order, depending on the length of the series. We allow the number of variables to increase to infinity by letting the maximal autogressive order increase as the length of the time series increases. Using an adaptation of a theorem in Portnoy (1985), we obtain an upper bound for this rate of increase such that the FIC theory still holds. The aim is to predict the series $\{y_t\}$, based on an AR-model estimated from $\{x_t\}$.

At this point, we introduce some notation for the "direct" model (3.3). First, denote the vectors $\boldsymbol{x}_t(p,h) = (x_{t-h}, \ldots, x_{t+1-h-p})^t$, $\boldsymbol{y}(p) = (y_T, \ldots, y_{T+1-p})^t$, and $\boldsymbol{\phi}(p,h) = \big(\phi_1(p,h), \ldots, \phi_p(p,h)\big)^t$. The OLS-estimates based on the series $\{x_t\}$ of the parameters $\boldsymbol{\phi}(p,h)$ are $\hat{\boldsymbol{\phi}}(p,h)$. Consequently, the h-step ahead prediction of the series $\{y_t\}$ is $\hat{y}_{T+h} = \hat{\boldsymbol{\phi}}(p,h)^t \boldsymbol{y}(p)$ if $1 \leq p \leq p_T$, and $\hat{y}_{T+h} = 0$ for $p = 0$. Because our goal is to make this prediction as accurate as possible, we take as focus parameter $\mu(p,h) = \boldsymbol{\phi}(p,h)^t \boldsymbol{y}(p)$.

Our goal is now to construct an information criterion aimed at selecting the model yielding the "best" estimate for the focus parameter from the $p_T + 1$ possible AR(p)-models. "Best" is defined in the sense of having the lowest mean squared forecast error. If we select the order p too low, the h-step ahead prediction of the series $\{y_t\}$ will be biased. On the other hand, choosing p too high will inflate the variance of the prediction. Therefore, we need to select p such that the h-step ahead prediction has at the same time a small bias and a small variance.

To define the Focussed Information Criterion, we assume the same setting as

in Claeskens and Hjort (2003). In particular, the results for the FIC apply in a local misspecification setting where the true, or optimal, values of the focus parameters are $\mu_{\text{true}} = \boldsymbol{\delta}(p_T)^t \boldsymbol{y}(p_T) T^{-1/2}$. The vector $\boldsymbol{\delta}$ is a fixed (though unknown) vector of infinite length, of which for practical purposes the first p_T components are used, which are denoted by $\boldsymbol{\delta}(p_T)$. A similar local misspecification setup is assumed (see Le Cam and Yang, 1990) for Le Cam's contiguity results, and local asymptotic normality, and in calculations under local alternatives for hypothesis testing problems. Let $\boldsymbol{J}_{T,\text{full}}$ be the estimated $p_T \times p_T$ information matrix of the AR(p_T)-model, the largest model under consideration, and assume that this matrix is of full rank. Since we use straightforward OLS-estimation for the parameters, this matrix can be estimated by

$$\hat{\boldsymbol{J}}_{T,\text{full}} = \frac{\hat{\boldsymbol{R}}(p_T, h)}{\hat{\sigma}^2(p_T, h)}.$$

Here

$$\hat{\boldsymbol{R}}(p_T, h) = \frac{1}{T + 1 - h - p_T} \sum_{t=p_T+h}^{T} \boldsymbol{x}_t(p_T, h) \boldsymbol{x}_t(p_T, h)^t \tag{3.5}$$

is the estimated autocovariance matrix of order p_T of the series $\{x_t\}$, and $\hat{\sigma}^2(p_T, h)$ is the estimated variance of the residuals after OLS-estimation. The matrix $\boldsymbol{R}(p_T)$ is the true autocovariance matrix of order p_T, and $\sigma^2(p_T, h)$ the true variance of the error terms. Using the ML-estimator would increase the complexity of the information matrix in the finite sample setting, while OLS and ML lead to the same limit expression for $\boldsymbol{J}_{T,\text{full}}$. We define the matrices $\hat{\boldsymbol{K}}_{T,p} = \hat{\sigma}^2(p_T, h) \hat{\boldsymbol{R}}(p, h)^{-1}$, and

$$\hat{\boldsymbol{M}}_{T,p} = \hat{\sigma}^2(p_T, h) \begin{pmatrix} \hat{\boldsymbol{R}}(p, h)^{-1} & 0 \\ 0 & 0 \end{pmatrix} \text{ of dimension } p_T \times p_T.$$

Finally, define

$$\boldsymbol{D}_T = \hat{\boldsymbol{\delta}}(p_T, h) = \sqrt{T} \hat{\boldsymbol{\phi}}(p_T, h).$$

The following proposition states the limit distribution of the estimated focus parameter. This result is the cornerstone of the Focussed Information Criterion when applied in this setting. The proof is found in Appendix A.2. Using similar

notation as in Claeskens and Hjort (2003), we set

$$\boldsymbol{\delta}(p_T, h) = \sqrt{T}\boldsymbol{\phi}(p_T, h)$$

and

$$\boldsymbol{\delta}(p, h) = \begin{pmatrix} \boldsymbol{R}(p,h)^{-1} & 0 \\ 0 & 0 \end{pmatrix} \boldsymbol{R}(p_T, h)\boldsymbol{\delta}(p_T, h).$$

Proposition 3.1 *Take h fixed and let $\hat{\mu}(p,h) = \hat{\phi}(p,h)'y(p)$ be the h-step ahead forecast of the true value μ_{true}. Under conditions (A1), (A2), (A3) listed in Appendix A.2, and if*

$$\frac{p_T\sqrt{\log T}}{T} \to 0 \text{ as } T \to \infty,$$

then we have, for every $0 \le p \le p_T$,

$$\sqrt{T}(\hat{\mu}(p,h) - \mu_{\text{true}}) \xrightarrow{d} \Lambda_p, \text{ for } T \to \infty, \tag{3.6}$$

where Λ_p is normally distributed with mean and variance given by

$$\lambda_p = \mathrm{E}[\Lambda_p] = \lim_{T\to\infty} \boldsymbol{y}(p_T)^t\big(\boldsymbol{\delta}(p,h) - \boldsymbol{\delta}(p_T,h)\big) \tag{3.7}$$

$$\sigma_p^2 = \mathrm{Var}(\Lambda_p) = \boldsymbol{y}(p)^t\boldsymbol{R}(p,h)^{-1}\boldsymbol{y}(p)\lim_{T\to\infty}\sigma^2(p_T,h). \tag{3.8}$$

This proposition does not assume that the time series $\{x_t\}$ and $\{y_t\}$ are independent. In fact, the results remain valid for $y_t = x_t$, stating the proposition for the single-series setting, but conditional on the observed data.

Hjort and Claeskens (2003) prove (although not specifically for time series) that the proposition holds for a finite maximal AR-order p_T. The additional condition on the rate of increase of p_T is a result of an adaptation of Theorem 3.2 in Portnoy (1985), which is formulated as Lemma A.1 in Appendix A.2, where the proof of Proposition 3.1 may also be found. The distribution of Λ_p in (3.6) is normal, with non-zero mean due to the local misspecification setting in which we work.

The distribution of Λ_p is the key result upon which the FIC is constructed. Specifically, the limiting distribution has mean squared error

$$\begin{aligned} r(p) &= \lim_{T\to\infty} \boldsymbol{y}(p_T)^t\big(\boldsymbol{\delta}(p,h)-\boldsymbol{\delta}(p_T,h)\big)\big(\boldsymbol{\delta}(p,h)-\boldsymbol{\delta}(p_T,h)\big)^t\boldsymbol{y}(p_T) \\ &\quad + \boldsymbol{y}(p)^t\boldsymbol{R}(p,h)^{-1}\boldsymbol{y}(p)\lim_{T\to\infty}\sigma^2(p_T,h). \end{aligned}$$

The FIC estimates this risk quantity for each AR-order p under consideration. To estimate $r(p)$, we estimate the unknown $\boldsymbol{R}(p_T,h)$ and $\sigma^2(p_T,h)$ by $\hat{\boldsymbol{R}}(p_T,h)$, see (3.5), and $\hat{\sigma}^2(p_T,h)$. We also unbiasedly estimate the quantity $\boldsymbol{\delta}(p_T,h)\boldsymbol{\delta}(p_T,h)^t$ by $\hat{\boldsymbol{\delta}}(p_T,h)\hat{\boldsymbol{\delta}}(p_T,h)^t-\hat{\sigma}^2(p_T,h)\hat{\boldsymbol{R}}(p_T,h)^{-1}$, where we calculated the covariance of the estimated parameters as $\mathrm{Cov}\big(\hat{\boldsymbol{\delta}}(p_T,h)\big)=\sigma^2(p_T,h)\boldsymbol{R}(p_T,h)^{-1}$. Finally, we drop the limit of T tending to infinity. After some algebraic manipulation, we get

$$\begin{aligned} \hat{r}(p) &= \Big(\boldsymbol{y}(p_T)^t\big(\hat{\boldsymbol{\delta}}(p,h)-\hat{\boldsymbol{\delta}}(p_T,h)\big)\Big)^2+2\hat{\sigma}^2(p_T,h)\boldsymbol{y}(p)^t\hat{\boldsymbol{R}}(p,h)^{-1}\boldsymbol{y}(p) \\ &\quad -\hat{\sigma}^2(p_T,h)\boldsymbol{y}(p_T)^t\hat{\boldsymbol{R}}(p_T,h)^{-1}\boldsymbol{y}(p_T). \end{aligned}$$

If we add $\hat{\sigma}^2(p_T,h)\boldsymbol{y}(p_T)^t\hat{\boldsymbol{R}}(p_T,h)^{-1}\boldsymbol{y}(p_T)$, which is independent of p, we arrive at the more compact expression for the FIC:

$$\mathrm{FIC}_p=\Big(\boldsymbol{y}(p_T)^t\big(\hat{\boldsymbol{\delta}}(p,h)-\hat{\boldsymbol{\delta}}(p_T,h)\big)\Big)^2+2\hat{\sigma}^2(p_T,h)\boldsymbol{y}(p)^t\hat{\boldsymbol{R}}(p,h)^{-1}\boldsymbol{y}(p). \quad (3.9)$$

We select the AR-order p with the smallest value for the FIC_p.

3.4 Simulations

We present the results of a simulation study to examine the performance of FIC compared to AIC and BIC, both in the one-series setting and in the two-series setting. Recall that, in Section 3.3, we estimated the parameters and selected the AR-order using one series $\{x_t\}$, and assumed that the actual prediction is done on a different series $\{y_t\}$, independent of $\{x_t\}$, though with the same stochastic structure. This is a similar setup as in Shibata (1980), Bhansali (1996) and Lee and Karagrigoriou (2001). In practical applications, however, such a situation does not often occur. Instead there is only a single time series available, and

model selection, as well as parameter estimation and prediction, have to be done using this single time series.

We performed simulation experiments to compare the performance of FIC with the classical AIC (Akaike, 1974) and BIC (Schwarz, 1978). These two establised criteria are defined as

$$-2\ell(x_t, \boldsymbol{\phi}) + C(T)p,$$

where $\ell(x_t, \boldsymbol{\phi})$ is the log-likelihood of the time series $\{x_t\}$, $C(T)$ a constant depending only on the length of the series, and p the AR-order of the considered model. For AIC, we have that $C(T)$ equals 2, and for BIC, we have $C(T) = \log(T)$. In all our studies, the true data-generating process is an ARMA(1,1)-model

$$Z_t = \phi Z_{t-1} + \varepsilon_t + \eta\varepsilon_{t-1},$$

where $\varepsilon_t \sim \mathcal{N}(0,1)$ i.i.d., and both ϕ and η take values in $\{-0.9, -0.7, \ldots, 0.9\}$. The stationarity and invertibility conditions on the parameters in this model reduce to $|\phi| < 1$ and $|\eta| < 1$. Hence, the ARMA(1,1)-model has an AR(∞)-representation. We let both parameters ϕ and η vary to examine whether or not the relative performance of the different information criteria depends on the values of these parameters. Note that, although the true data-generating process is an ARMA(1,1)-model, this model is not included in the group of considered models, which are all autoregressive models of finite order. Hence, the selected model will always be the "best approximating" model among the candidate autoregressive models.

In the first simulation experiment, we generate for each setting $M = 10\,000$ series $\{x_t\}$ of length $T = 200$, which we use for both model order selection and parameter estimation, and on which we will construct our predictions. This series $\{x_t\}$ is generated up to length $T + h$ to allow an out-of-sample estimate of the prediction accuracy of the h-step ahead forecast of $\{x_t\}$. We select a model as in (3.3) for $0 \leq p \leq p_T$, and $h = 2$, yielding the "best" finite-order AR approximation of the series $\{x_t\}$. We have chosen the maximal order $p_T = 20 = \sqrt{2T}$ such that a sufficient, but not excessive, number of models is considered. In practice, the choice of the maximal AR-order is somewhat arbitrary, and will be

increased by one if the largest model was chosen, to account for the possibility of long-term dependencies in the time series. For each simulation run, the selection is done by AIC, BIC and FIC. Once the model is selected, a h-step ahead forecast is made of this series $\{x_t\}$ using the estimated parameters of the selected model. This forecast is denoted by $\hat{x}^{(j)}_{T+h} = \boldsymbol{x}(p_{h,j})^t \hat{\boldsymbol{\phi}}(p_{h,j}, h)$, where j is the number of the simulation run, and $p_{h,j}$ is the AR-order of the model selected for the h-step ahead forecast in simulation run j.

For each simulation setting in the experiment above, we present the MSE of the h-step ahead prediction of the series $\{y_t\}$, where the prediction is performed using the models selected by (i) AIC, (ii) BIC, and (iii) FIC. We define the MSE by

$$\mathrm{MSE}(\hat{x}_{T+h}) = \frac{1}{M}\sum_{j=1}^{M}(\hat{x}^{(j)}_{T+h} - x^{(j)}_{T+h})^2,$$

with $\hat{x}^{(j)}_{T+h}$ as defined above, and with $x^{(j)}_{T+h}$ the true generated value of the series $\{x_t\}$ in the j-th simulation. We also define the Relative Mean Squared Error as

$$\mathrm{rMSE}(\hat{x}_{T+h}, \mathrm{xIC}_1, \mathrm{xIC}_2) = \frac{\mathrm{MSE}(\hat{x}_{T+h,\mathrm{xIC}_1})}{\mathrm{MSE}(\hat{x}_{T+h,\mathrm{xIC}_2})}, \tag{3.10}$$

where $\hat{x}_{T+h,\mathrm{xIC}_1}$ and $\hat{x}_{T+h,\mathrm{xIC}_2}$ are the h-step ahead predictions of the series $\{x_t\}$ made with models chosen by respectively xIC_1 and xIC_2 as information criteria. When the relative MSE is smaller than 1, xIC_1 selects models with a lower MSE for the h-step ahead prediction than xIC_2.

Table 3.1 presents the simulated relative MSEs of the models selected by FIC with respect to those selected by the AIC (relative $\mathrm{MSE}(\hat{x}_{T+h}, \mathrm{FIC}, \mathrm{AIC})$, top tables), and those with respect to those selected by the BIC ($\mathrm{rMSE}(\hat{x}_{T+h}, \mathrm{FIC}, \mathrm{BIC})$, bottom tables). These tables show that the performances of AIC and BIC was slightly better (a few percent) than that of FIC. Note that this occured for all 100 different settings of parameters (ϕ, η). Standard errors for the MSE ratios have been computed via the delta method and are approximately 5×10^{-3}, this due to the large number of simulation runs. Hence we conclude that there is statistical evidence that AIC and BIC yield lower MSEs in this simulation experiment than FIC, but the practical difference in performance among the procedures remains small.

ϕ/η	-0.9	-0.7	-0.5	-0.3	-0.1	0.1	0.3	0.5	0.7	0.9
-0.9	1.023	1.045	1.044	1.042	1.048	1.047	1.045	1.041	1.033	1.044
-0.7	1.023	1.036	1.044	1.057	1.057	1.050	1.047	1.025	1.056	1.027
-0.5	1.025	1.028	1.038	1.052	1.049	1.045	1.033	1.049	1.033	1.014
-0.3	1.022	1.036	1.036	1.032	1.042	1.039	1.049	1.039	1.032	1.025
-0.1	1.036	1.033	1.043	1.040	1.031	1.037	1.040	1.038	1.038	1.039
0.1	1.038	1.044	1.041	1.046	1.044	1.053	1.028	1.031	1.038	1.030
0.3	1.020	1.041	1.047	1.047	1.045	1.031	1.040	1.037	1.036	1.026
0.5	1.026	1.035	1.049	1.050	1.041	1.052	1.045	1.044	1.034	1.026
0.7	1.014	1.040	1.034	1.043	1.052	1.047	1.054	1.045	1.039	1.025
0.9	1.040	1.044	1.052	1.038	1.054	1.047	1.042	1.040	1.042	1.020

ϕ/η	-0.9	-0.7	-0.5	-0.3	-0.1	0.1	0.3	0.5	0.7	0.9
-0.9	1.017	1.043	1.046	1.049	1.065	1.062	1.052	1.041	1.026	1.060
-0.7	1.025	1.040	1.043	1.071	1.077	1.065	1.055	1.030	1.073	1.039
-0.5	1.035	1.039	1.052	1.067	1.056	1.053	1.046	1.061	1.043	1.012
-0.3	1.025	1.045	1.040	1.041	1.057	1.053	1.063	1.054	1.044	1.037
-0.1	1.059	1.054	1.060	1.058	1.046	1.047	1.057	1.055	1.058	1.059
0.1	1.063	1.061	1.054	1.063	1.057	1.070	1.042	1.048	1.058	1.054
0.3	1.035	1.050	1.061	1.064	1.059	1.048	1.047	1.044	1.048	1.037
0.5	1.035	1.053	1.056	1.061	1.047	1.066	1.061	1.054	1.045	1.040
0.7	1.026	1.054	1.040	1.054	1.072	1.067	1.066	1.054	1.040	1.030
0.9	1.057	1.039	1.051	1.045	1.066	1.066	1.054	1.048	1.043	1.020

Table 3.1: Ratios of mean squared errors for the 2-step ahead prediction of the series $\{x_t\}$, with model order selection using the same series, and prediction according to the *direct method*. An ARMA(1,1)-process generated the series $\{x_t\}$. The autoregression parameter ϕ can be found in the leftmost column, and the moving average parameter η is indicated in the top row. The upper table shows the rMSE$(\cdot, \mathrm{FIC}, \mathrm{AIC})$, the lower table shows the rMSE$(\cdot, \mathrm{FIC}, \mathrm{BIC})$, as defined in (3.10).

By contrast, if we repeat the experiment with the maximum length of the series larger ($T = 500$ or $T = 2\,000$, results available upon request), we find that these ratios become even closer to 1, lending empirical support for the statement that FIC performs as well as AIC and BIC asymptotically. This is due to the fact that FIC is an unbiased estimator of the asymptotic MSE for the h-step ahead prediction, thereby leaving out of the MSE a constant term that does not depend on the model. In other words, we select the model with the smallest estimated mean squared forecast error (MSFE). Hence, it is expected that FIC asymptotically selects the model with the lowest MSFE. As a result, FIC and AIC will have the same asymptotic performance, as the asymptotic efficiency of AIC was proved by Ing and Wei (2005) in the single-series case.

Let us now discuss which models are selected by the three criteria. First, we examine the case where $\phi = \eta = -0.9$, which is far from a white noise process. In that case, AIC selected model orders ranging from 1 to $p_T = 20$, with the maximal order being selected only 38 times out of 10 000. In 59.5% of the cases AIC selected a model order between 4 and 7. As expected, BIC selected lower orders, with AR-order 9 being the maximum, and in 80.4% of the cases choosing 2 or 3 as the order. FIC selected on average an order somewhere between that of AIC and BIC, with 70% of the selected AR-orders between 1 and 5, and with the maximal order chosen 42 times out of 10 000. AIC and FIC seleced the same model order in about 11.5% of the cases, and FIC and BIC agreed on the selected model order in 19% of the cases.

Closer to the white noise case, $\phi = \eta = 0.1$ for example, AIC selected 0 or 1 as model order at least 3 times out of 4, with the white noise case chosen 70% of the time. The maximal order was selected only 4 times. BIC selected the white noise model in more than 96 cases out of 100. Finally, FIC selected the white noise case in 41% of the cases, while at least 75% of the selected orders were 5 or below. Nevertheless, the maximal model order $p_T = 20$ was chosen 61 times out of 10 000. Here, we see that AIC and FIC agreed in 36% of the cases, and that BIC and FIC agreed in 40% of the cases.

In Section 3.3 the FIC was derived from the setting where we have one time

series $\{x_t\}$ available for model selection and parameter estimation, and another stochastically independent time series $\{y_t\}$ for prediction. We have conducted a second simulation experiment to compare the two-series framework with the more realistic one-series framework. This simulation experiment was set up along the same lines as the one-series experiment, with the following difference. For each parameter setting (ϕ, η) we generated $M_x = 100$ different series $\{x_t\}$ for model selection and parameter estimation. Then, for each of these series, we generated $M_y = 100$ independent series $\{y_t\}$ which we will forecast. Since we wish to compare the performance of the different model selection criteria in MSE sense, we generated the series $\{y_t\}$ up to length $T + h$. After running the experiment for the two-series setting, we found similar results as for the one-series setting. Figure 3.1 shows the results of a comparison of the performance between the two-series and the one series setting, where the selection criterion used is the FIC. The surface shown in the figure depicts the relative MSE of the one-series FIC with respect to the two-series FIC. Where the surface is above 1, the grey-shaded facets, using two independent series resulted in better performance. It is obvious that both settings resulted in very similar performances, and that there was no clear preference for either setting. Indeed, the two-series setting was superior to the one-series setting for 58 of the 100 parameter choices.

3.5 Real data applications

In this section we compare the performances of AIC, BIC, and FIC on two real datasets: monthly US liquor sales data (Diebold, 2001, p. 54), and monthly life insurance data (data available at
http:// www.econ.kuleuven.be/public/ndbae06/courses/dynmodels/assvie.xls).
The life insurance dataset goes from January 1964 to December 1980, and denotes the net number of new personal life insurances for a large insurance company. Since the theory above is developed for stationary series, we first removed the trend and seasonality effects. First, we took the logarithm of the series to make the variance of the innovation terms constant over time. Next, we took the first differences to remove the trend, and take seasonal differences to remove the

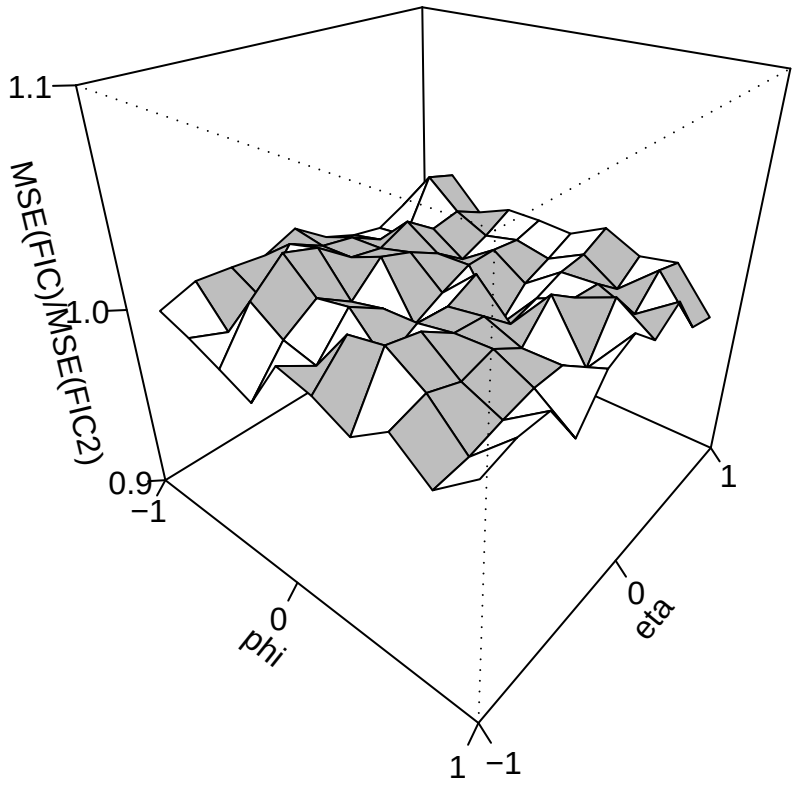

Figure 3.1: 3D-surface plot for the ratios of mean squared errors for the 2-step ahead prediction of the series $\{x_t\}$, comparing model order selection using the series $\{x_t\}$ with model order selection using the series $\{y_t\}$, and where prediction is according to the *direct method*. An ARMA(1,1)-process generated both series $\{x_t\}$ and $\{y_t\}$. The autoregression parameter ϕ can be found on the phi axis, and the moving average parameter η is indicated on the eta axis. The surface shows the ratios of MSEs where the selection criterion used in both cases is the FIC. Where this surface lies above 1, signified by the grey-shaded facets, the two-series case had a smaller MSE than the one-series case.

seasonality effects, so that we had a stationary series. Out-of-sample h-step ahead forecasting was used to estimate the mean squared errors for each of the three information criteria, this for horizons $h = 1, \ldots, 5$. More precisely, we started with the first half of the series $\{x_t\}$, that is $1 \leq t \leq T/2$, and predicted $x_{T/2+h}$. We then added the next observation, $x_{T/2+1}$, and based on $\{x_t\}$, $1 \leq t \leq T/2+1$, predicted $x_{T/2+1+h}$. This process was repeated until we had used all observations up to and including x_{T-h} to predict x_T. Note that the order of the selected model depends on the time index t at which the prediction for x_{t+h} is made. We chose the maximal AR-orders of the models equal to $p_T = 15$. The maximal order was chosen to be approximately equal to $\sqrt{T}$, such that a sufficient but not excessive number of models is considered. Next, we performed a pairwise comparison of the estimated MSEs for each h, and tested whether there are significant differences. The MSEs are estimated as

$$\text{MSE} = \frac{1}{T/2+1} \sum_{t=T/2}^{T-h} (x_{t+h} - \hat{x}_{t+h})^2.$$

The pairwise comparison was done by the Diebold-Mariano test (Diebold, 2001, p. 293-294), which is basically a type of paired t-test for equality of means. In this case however, the data consisted of squared residuals, one group for each information criterion. As it is likely that there is serial correlation in these residuals, special care had to be taken to determine the standard error used in computing the t-values.

Table 3.2 shows the estimated mean squared errors for the different prediction horizons h and the different order selection criteria, together with the average value of the selected orders of the autoregressive model. It also shows the t-values and corresponding p-values for the Diebold-Mariano tests. The results reported here are valid when the plug-in method for prediction is used. We repeated the experiment with the direct method for prediction and we did not find a significant difference with the plug-in method. The upper table shows the resulting values for the US liquor sales time series, and the bottom table shows the corresponding results for the Life Insurance time series. A positive t-value means that the first criterion leads to predictions with a higher MSE than the second criterion.

(a)	$h = 1$	$h = 2$	$h = 3$	$h = 4$	$h = 5$
MSE(AIC)	1.153 (12.70)	1.516 (12.70)	1.509 (12.70)	1.566 (12.70)	1.630 (12.70)
MSE(BIC)	1.392 (2.00)	1.715 (2.00)	1.713 (2.00)	1.781 (2.00)	1.855 (2.00)
MSE(FIC)	1.176 (4.59)	1.504 (4.72)	1.528 (4.82)	1.591 (5.16)	1.666 (4.70)
MSE(MSE)	1.140 (14.00)	1.486 (15.00)	1.490 (15.00)	1.544 (15.00)	1.610 (14.00)
			Diebold-Mariano test results		
AIC–FIC	–0.818 (0.413)	0.312 (0.755)	–0.314 (0.753)	–0.740 (0.459)	–0.549 (0.583)
BIC–FIC	2.971 (0.003)	4.003 (0.000)	2.696 (0.007)	2.545 (0.011)	1.931 (0.053)
MSE–FIC	–1.322 (0.186)	–0.438 (0.661)	–0.639 (0.523)	–1.455 (0.146)	–0.897 (0.369)

(b)	$h = 1$	$h = 2$	$h = 3$	$h = 4$	$h = 5$
MSE(AIC)	94.51 (7.04)	149.46 (7.04)	134.68 (7.04)	120.46 (7.04)	113.94 (7.04)
MSE(BIC)	76.77 (3.00)	119.71 (3.00)	117.13 (3.00)	120.32 (3.00)	118.06 (3.00)
MSE(FIC)	84.74 (3.31)	137.28 (4.11)	124.18 (4.39)	118.36 (4.30)	120.91 (4.41)
MSE(MSE)	76.56 (5.00)	115.30 (0.00)	115.83 (6.00)	117.07 (12.00)	114.95 (12.00)
			Diebold-Mariano test results		
AIC–FIC	1.892 (0.059)	1.300 (0.194)	1.006 (0.314)	0.180 (0.857)	–0.560 (0.575)
BIC–FIC	–0.807 (0.420)	–1.191 (0.234)	–0.549 (0.583)	0.142 (0.887)	–0.227 (0.820)
MSE–FIC	–0.756 (0.450)	–1.643 (0.100)	–0.694 (0.488)	–0.137 (0.891)	–0.877 (0.380)

Table 3.2: Comparison of models selected by the information criteria FIC, AIC, and BIC. A further comparison is made with a model selected based on the MSE of a hold-out sample. The table contains the estimated mean squared errors ($\times 10^{-3}$) for each prediction horizon h, with the average value of the selected order within parenthesis. Furthermore, t-values (p-values) of the Diebold-Mariano test for pairwise differences in MSE are presented. Results are given in (a) for the US Liquor sales data, and in (b) for the life insurance data.

For the US liquor sales time series, we observed that there were no significant differences in performance between AIC and FIC. On the other hand, the BIC performed significantly worse than both AIC and FIC. For the Life Insurance time series, FIC performed slightly, but not significantly, better than AIC. The BIC performed slightly better than the FIC on these data, but again, the difference was not significant. To conclude, the Diebold-Mariano test showed that the three different information criteria performed about equally well on the two examples considered, hereby confirming the results of Section 3.4. The power of the Diebold-Mariano test might be too low to detect differences between forecast methods, especially in small-sample settings (Harvey, Leybourne and Newbold, 1997 and 1998). However, we have also compared the forecast methods using the modified Morgan-Granger-Newbold test (Harvey, Leybourne and Newbold, 1997) and arrived at the same conclusions as with the Diebold-Mariano test. A further comparison can be made with selection based on the MSE of a hold-out sample. Since we directly minimised the value of the MSE, we expected this criterion to perform very well. Indeed, for both datasets we observed that directly selecting the model order which minimises MSE yielded more accurate estimates than when FIC was used, although the differences were not significant according to the Diebold-Mariano test (p-values > 0.1).

Increasing the maximal order to $p_T = 50$, rather than taking $p_T = 15 \approx \sqrt{T}$, had negligible influence on the performance of the criteria AIC, BIC and FIC. Choosing the order too small, say $p_T = 5$, mostly affected AIC and FIC, because BIC has a natural tendency to select simple models.

Performances of the criteria were also compared using the Mean Absolute Error (MAE) and the Mean Absolute Percentage Error (MAPE)

$$\text{MAE} = \frac{1}{T/2+1}\sum_{t=T/2}^{T-h}|x_{t+h} - \hat{x}_{t+h}| \text{ and MAPE} = \frac{1}{T/2+1}\sum_{t=T/2}^{T-h}\Big|\frac{x_{t+h} - \hat{x}_{t+h}}{x_{t+h}}\Big|.$$

Using MAE as the loss function in the Diebold-Mariano test gave similar results as in Table 3.2. However, using MAPE as loss function, there was a preference for FIC for all prediction horizons, though not significant. The Diebold-Mariano test had p-values approximately 0.25 (0.30) for the comparison with AIC (BIC).

3.6 Extensions

In this section we list three extensions of the main ideas in this paper. First we explain how the results need to be adapted for prediction with the plug-in method. Second, we provide the expression for FIC when the impulse response is the focus parameter. Third, we obtain a definition of FIC for simultaneous selection of regression variables and the autoregressive order of the error terms.

3.6.1 Using plug-in methods

Direct prediction results in a h-step ahead predictor which is a linear combination of the parameter estimates. Therefore Proposition 3.1 is applicable. By contrast, the plug-in method leads to a predictor which is a polynomial of order h of the parameter estimates (see equation (3.4)). In order to derive the distribution of the predictor in each candidate model, the first main step is to show that a suitably scaled version of $\sqrt{T}\Big(\boldsymbol{g}\big(\hat{\boldsymbol{\phi}}(p_T)\big) - \boldsymbol{g}\big(\boldsymbol{\phi}_{\text{true}}(p_T)\big)\Big)^t \boldsymbol{y}(p_T)$ has an asymptotic normal distribution denoted Λ_{p_T}, where $\boldsymbol{g}\big(\hat{\boldsymbol{\phi}}(p_T)\big) = \Big(g_1\big(\hat{\boldsymbol{\phi}}(p_T)\big), \ldots, g_{p_T}\big(\hat{\boldsymbol{\phi}}(p_T)\big)\Big)^t$ with $g_i\big(\hat{\boldsymbol{\phi}}(p_T)\big)$ a polynomial of degree h in $\hat{\phi}_1(p_T), \ldots, \hat{\phi}_{p_T}(p_T)$. The argument in Appendix A.2 shows why this is the case in our setting. We then proceed by computing the limiting mean squared error of Λ_{p_T}, and by estimating this quantity in an unbiased way. This estimator is the FIC, which is then computed for each candidate autoregressive order p. In our setting, it has the same form as the FIC for the direct method (3.9), but with $\boldsymbol{y}(p_T)$ replaced by the recursively defined

$$\hat{\boldsymbol{\omega}}_h(p_T) = \hat{\boldsymbol{m}}_h(p_T) + \hat{\boldsymbol{\Omega}}_h(p_T)\hat{\boldsymbol{\phi}}(p_T).$$

Here $\hat{\boldsymbol{m}}_h(p_T) = (\hat{y}_{T+h-1}(p_T), \ldots, \hat{y}_{T+h-p_T}(p_T))^t$ with $\hat{y}_{T+i}$ defined as in (3.4). Also, $\hat{\boldsymbol{\Omega}}_h(p_T) = (\hat{\boldsymbol{\omega}}_{h-1}(p_T), \ldots, \hat{\boldsymbol{\omega}}_{h-p_T}(p_T))$ where $\hat{\boldsymbol{\omega}}_i(p_T) = \boldsymbol{0}$ for $i \leq 0$. The $\boldsymbol{y}(p)$ in expression (3.9) are replaced by a vector containing the first p elements of $\hat{\boldsymbol{\omega}}_h(p_T)$. This yields the FIC we have used in the simulations of Section 3.4 for the plug-in method for prediction. The model selected is, as before, the model with the lowest value of FIC.

In a simulation experiment, we compared the direct and the plug-in method

for prediction, simarly as in Section 4. The conclusions for the plug-in method were the same as those for the direct method: AIC and BIC were slightly better than FIC (just a few percent in MSE sense, though this difference is significant). Of interest also is whether the direct method and plug-in method for prediction are equivalent. This can be seen in Figure 3.2. This figure shows the relative MSEs, where we compare the plug-in method with the direct method, for model selection with the FIC. A value larger than 1, the grey shaded facets, indicates that the direct method results in a forecast with lower MSE. As we can see, there was no clear preference for either method, as both prediction methods came out as best for roughly half of the settings.

3.6.2 Focus on the impulse response

Up to now, the goal was to select the autoregressive order p with which to obtain the h-step ahead predictor with the smallest value of the FIC. Here we change focus to the impulse response at lag τ, denoted $\imath(\tau)$. The impulse response function $\imath(\tau)$ is equal to a time series that is the realization of the data-generating process for which all innovation terms ε_t are set equal to zero, except for $\varepsilon_0 = 1$ (see Hamilton 1994, page 5). The impulse response function is often used by economists to study the effect of innovation shocks to the variable of interest. Here we want to use the FIC to select the best AR-order for making estimates of the impulse response function at a certain lag. This problem has already been investigated via a simulation study in Hansen (2005). Here we give a theoretical justification for the use of the FIC in this setting.

We use the same notation as in Section 3.2. The focus parameter μ introduced in Section 3.3 is replaced by $\mu = \imath(\tau)$. The plug-in method based on model (3.1) leads to the following estimated focus parameter:

$$\hat{\mu} = \hat{\imath}(\tau) = \hat{\phi}_1(p)\hat{\imath}(\tau - 1) + \cdots + \hat{\phi}_p(p)\hat{\imath}(\tau - p),$$

where $\hat{\imath}(\tau) = 0$ for $\tau < 0$ and $\hat{\imath}(0) = 1$. From this expression it is clear that estimating the impulse response of a time series at lag τ is a special case of a τ-step ahead prediction, applied to a time series with 0 on every time t, except

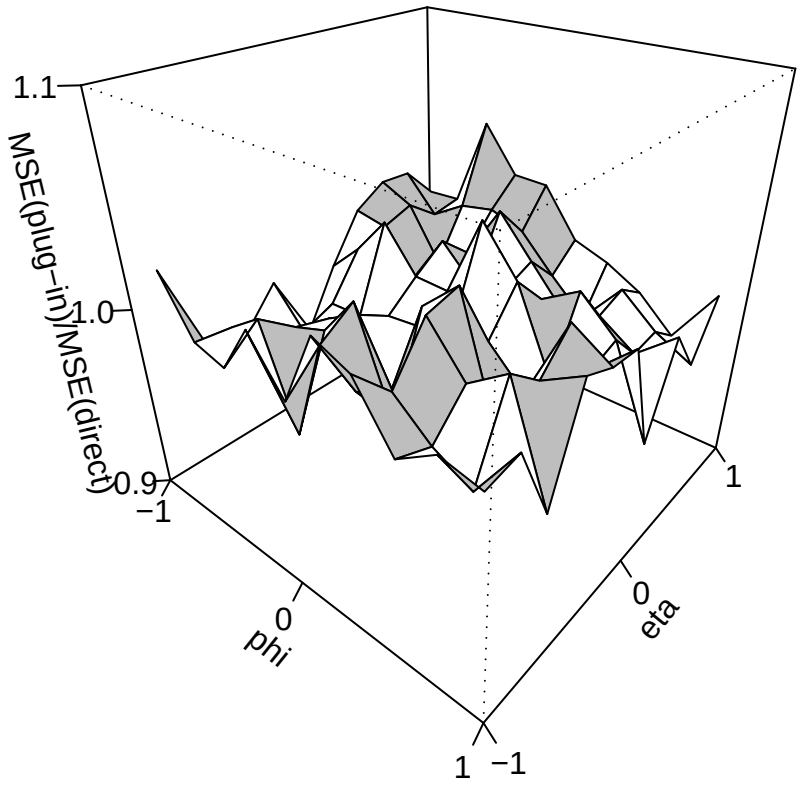

Figure 3.2: 3D-surface plot for the ratios of mean squared errors for the 2-step ahead prediction of the series $\{x_t\}$, with model order selection using the series $\{x_t\}$, comparing prediction with the plug-in method and with the direct method. An ARMA(1,1)-process generated the series $\{x_t\}$. The autoregression parameter ϕ can be found on the phi axis, and the moving average parameter η is indicated on the eta axis. The surface shows the ratios of MSEs where the selection criterion used in both cases is the FIC. Where this surface lies above 1, signified by the grey-shaded facets, the direct method for prediction resulted in a lower MSE than the plug-in method.

for a 1 on time T, where the parameter estimators are constructed from the given time series $\{x_t\}$. With this observation, the results of Proposition 3.1 are readily applicable for the impulse response as a focus parameter. From Section 3.6.1, it then follows that the FIC is an unbiased estimator of the limiting mean squared error of the impulse response in the case of a growing number of parameters. The expression for FIC_p for impulse response is then given as in the previous subsection for $h = \tau$, although now with $y_T = 1$ and $y_t = 0$ for $1 \leq t < T$.

To illustrate the use of the FIC for model selection where the focus is the impulse response function $\imath(\tau)$ at lag $\tau = 2$, we present the results of a simulation study similar to the one in Section 3.4. Here, we took the number of simulation runs for each setting as $M = 1\,000$, and we allowed the parameters of the simulated ARMA(1,1)-model to be in the range $(-0.9, -0.8, \ldots, 0.9)$. The results of this simulation are presented in Figure 3.3. This figure shows the relative MSE, $\text{rMSE}(\cdot, \text{AIC}, \text{FIC})$ as in (3.10), of the estimated impulse response function at lag $\tau = 2$. Where this surface lies above 1, corresponding to the grey-shaded facets, the FIC selected models with a lower MSE than the AIC. We observe that there are regions in the parameter space (ϕ, η) where the FIC performed significantly better than the AIC. In particular, when the series was close to a white noise ($|\eta + \phi| = 0$), there were pronounced differences. At present, we do not have a clear explanation for this.

3.6.3 Simultaneous selection of regression variables and the AR order

Up to now we considered stationary time series with zero mean. We implicitly assumed that the trend and the seasonality effects of this series were removed beforehand. We also ignored the possibility that there might be exogenous variables upon which the time series has been regressed prior to analysis. This is a commonly used approach when estimating and predicting time series: first identify and fit the deterministic component, and then determine the error-structure. However, if the identification of the deterministic component includes a variable selection step, Golan et al. (1996, Chapter 10) illustrated that the classical vari-

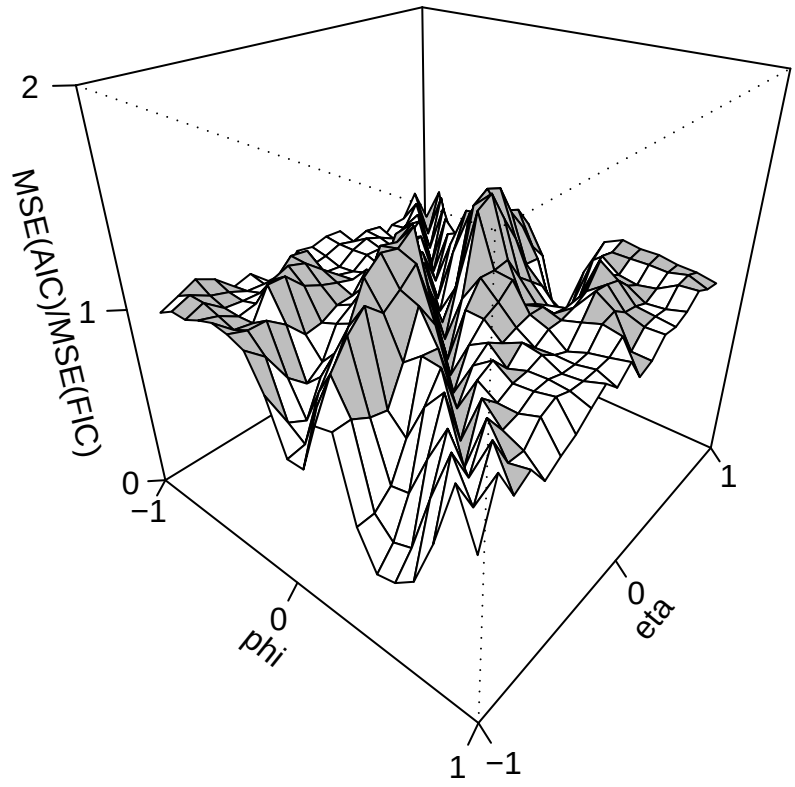

Figure 3.3: 3D-surface plot for the ratios of mean squared errors for the estimation of the impulse response function of the series $\{x_t\}$ at lag 2, with model order selection using the same series. An ARMA(1,1)-process generated the series $\{x_t\}$. The autoregression parameter ϕ can be found on the phi axis, and the moving average parameter η is indicated on the eta axis. The surface shows the ratios of MSEs where the AIC is compared with the FIC. Where this surface lies above 1, signified by the grey-shaded facets, the FIC selected models which results in a lower MSE than the AIC.

able selection criteria perform poorly if the residual errors do not satisfy the uncorrelatedness assumption. Recently, Shi and Tsai (2004) proposed an alternative selection criterion which simultaneously selects the regression variables for inclusion in the model and the autoregressive order of the error terms.

In a similar spirit, we can employ the FIC to perform simultaneous selection of the regression variables and of the AR-order of the model errors. Assume that we have a time series $\{y_t\}$ and explanatory series $\boldsymbol{x}_t = \big(\{x_{t,1}\}, \ldots, \{x_{t,k}\}\big)^t$, and that the data are generated from the following model

$$y_t = \boldsymbol{x}_t^t \boldsymbol{\beta} + u_t \quad \text{with} \quad u_t = \phi_1 u_{t-1} + \cdots + \phi_P u_{t-P} + \varepsilon_t, \tag{3.11}$$

where the errors ε_t are independent and identically normally distributed with mean 0 and variance σ^2 for $t = P+1, \ldots, T$, and where $\boldsymbol{U}_P = (u_1, \ldots, u_P)^t$ is distributed as $\mathcal{N}(0, \sigma^2 \boldsymbol{R}(P))$. The log-likelihood function under model (3.11) is then (omitting constants not depending on the model)

$$\ell(\boldsymbol{\beta}, \boldsymbol{\Phi}, \sigma^2) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2}\log|\boldsymbol{R}(P)| - \frac{1}{2\sigma^2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^t \boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}),$$

where $\boldsymbol{\Phi} = (\phi_1, \ldots, \phi_P)^t$, $\boldsymbol{Y} = (y_1, \ldots, y_T)^t$, $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)^t$, $\boldsymbol{U} = (u_1, \ldots, u_T)^t$, and $\boldsymbol{\Sigma} = \mathrm{Cov}(\boldsymbol{U})/\sigma^2$. Note that $\boldsymbol{\Sigma}$ and $\boldsymbol{R}(P)$ depend on $\boldsymbol{\Phi}$. The expressions for $|\boldsymbol{R}(P)|$ and $\boldsymbol{\Sigma}^{-1}$ can be found in Ljung and Box (1979). To facilitate the derivations, we condition on the first P observations, and write the conditional log-likelihood function as

$$\begin{aligned}
&\ell(\boldsymbol{\beta}, \boldsymbol{\Phi}, \sigma^2 \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_P, y_1, \ldots, y_P) \\
&\quad = -\frac{n}{2}\log\sigma^2 - \frac{1}{2}\log|\boldsymbol{R}(P)| - \frac{1}{2\sigma^2}\sum_{t=P+1}^{T}\Big(y_t - \boldsymbol{x}_t^t\boldsymbol{\beta} - \sum_{i=1}^{P}\phi_i(y_{t-i} - \boldsymbol{x}_{t-i}^t\boldsymbol{\beta})\Big)^2,
\end{aligned} \tag{3.12}$$

From this expression, we derive the estimated $(k+P)\times(k+P)$ information matrix $\boldsymbol{J}_{T,\text{full}}$. This matrix has components

$$\begin{aligned}
\big(\boldsymbol{J}_{T,\text{full}}\big)_{\beta_i,\beta_j} &= -\frac{1}{T-P}\cdot\frac{\partial^2\ell(\cdot)}{\partial\beta_i\partial\beta_j}\bigg|_{\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\Phi}},\hat{\sigma}^2} \\
&= \frac{1}{(T-P)\hat{\sigma}^2}\sum_{t=P+1}^{T}\Big(x_{t,i} - \sum_{l=1}^{P}\hat{\phi}_l x_{t-l,i}\Big)\Big(x_{t,j} - \sum_{l=1}^{P}\hat{\phi}_l x_{t-l,j}\Big),
\end{aligned}$$

$$\begin{aligned}\big(\boldsymbol{J}_{T,\text{full}}\big)_{\phi_i,\phi_j} &= -\frac{1}{T-P}\cdot\frac{\partial^2\ell(\cdot)}{\partial\phi_i\partial\phi_j}\Big|_{\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\Phi}},\hat{\sigma}^2}\\ &= \frac{1}{T-P}\cdot\frac{\partial^2\log|\boldsymbol{R}(P)|}{\partial\phi_i\partial\phi_j}\Big|_{\hat{\boldsymbol{\Phi}}}\\ &\qquad+\frac{1}{(T-P)\hat{\sigma}^2}\sum_{t=P+1}^{T}(y_{t-i}-\boldsymbol{x}_{t-i}^t\hat{\boldsymbol{\beta}})(y_{t-j}-\boldsymbol{x}_{t-j}^t\hat{\boldsymbol{\beta}}),\ \text{and}\end{aligned}$$

$$\begin{aligned}\big(\boldsymbol{J}_{T,\text{full}}\big)_{\beta_i,\phi_j} &= -\frac{1}{T-P}\cdot\frac{\partial^2\ell(\cdot)}{\partial\beta_i\partial\phi_j}\Big|_{\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\Phi}},\hat{\sigma}^2}\\ &= \frac{1}{(T-P)\hat{\sigma}^2}\sum_{t=P+1}^{n}\Big(x_{t-j,i}\Big(y_t-\boldsymbol{x}_l^t\hat{\boldsymbol{\beta}}-\sum_{l=1}^{P}\phi_l(y_{t-l}-\boldsymbol{x}_{t-l}^t\hat{\boldsymbol{\beta}})\Big)\\ &\qquad+(y_{t-j}-\boldsymbol{x}_{t-j}^t\hat{\boldsymbol{\beta}})\Big(x_{t,i}-\sum_{l=1}^{P}\hat{\phi}_l x_{t-l,i}\Big)\Big).\end{aligned}$$

For S a subset of $\{1,\ldots,k\}$ and $0\le p\le P$, let $\boldsymbol{\pi}_{S,p}$ be a projection matrix of dimension $(|S|+p)\times(k+P)$ mapping any vector $\boldsymbol{\nu}=(\nu_{1,1},\ldots,\nu_{1,k},\nu_{2,1},\ldots,\nu_{2,P})^t$ onto $(\boldsymbol{\nu}_S^t,\nu_{2,1},\ldots,\nu_{2,p})^t$, where $\boldsymbol{\nu}_S$ has components $\nu_{1,i}$ with $i\in S$. Denote $\boldsymbol{K}_{T,S,p}=(\boldsymbol{\pi}_{S,p}\boldsymbol{J}_{T,\text{full}}\boldsymbol{\pi}_{S,p}^t)^{-1}$ and $\boldsymbol{M}_{T,S,p}=\boldsymbol{\pi}_{S,p}^t\boldsymbol{K}_{T,S,p}\boldsymbol{\pi}_{S,p}$. The focus parameter in the FIC is the h-step ahead forecast, using the plug-in method for prediction:

$$\mu(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\Phi}})=\boldsymbol{x}_{T+h}^t\hat{\boldsymbol{\beta}}+\hat{\phi}_1(\hat{y}_{T+h-1}-\boldsymbol{x}_{T+h-1}^t\hat{\boldsymbol{\beta}})+\cdots+\hat{\phi}_P(\hat{y}_{T+h-P}-\boldsymbol{x}_{T+h-P}^t\hat{\boldsymbol{\beta}}),$$

and denote by $\boldsymbol{\omega}$ the vector with components

$$\begin{aligned}\omega_{1,i} &= -\frac{\partial\mu(\boldsymbol{\beta},\boldsymbol{\Phi})}{\partial\beta_i}\Big|_{\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\Phi}}} = -x_{T+h,i}-\sum_{j=1}^{P}\hat{\phi}_j\Big(\frac{\partial\hat{y}_{T+h-j}}{\partial\beta_i}\Big|_{\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\Phi}}}-x_{T+h-j,i}\Big)\quad 1\le i\le k\\ \omega_{2,j} &= -\frac{\partial\mu(\boldsymbol{\beta},\boldsymbol{\Phi})}{\partial\phi_j}\Big|_{\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\Phi}}} = -(\hat{y}_{T+h-j}-\boldsymbol{x}_{T+h-j}^t\hat{\boldsymbol{\beta}})-\sum_{i=1}^{P}\hat{\phi}_i\frac{\partial\hat{y}_{T+h-i}}{\partial\phi_j}\Big|_{\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\Phi}}}\quad 1\le j\le P,\end{aligned}$$

where $\hat{y}_t=y_t$ and hence $\partial\hat{y}_t/\partial\beta_i=\partial\hat{y}_t/\partial\phi_j=0$ for $t\le T$.

Combining these ingredients leads to

$$\text{FIC}_{S,p}=\boldsymbol{\omega}^t(\boldsymbol{I}-\boldsymbol{M}_{T,S,p}\boldsymbol{J}_{T,\text{full}})\hat{\boldsymbol{\delta}}\hat{\boldsymbol{\delta}}^t(\boldsymbol{I}-\boldsymbol{J}_{T,\text{full}}\boldsymbol{M}_{T,S,p})\boldsymbol{\omega}+2\boldsymbol{\omega}^t\boldsymbol{M}_{T,S,p}\boldsymbol{\omega},\quad(3.13)$$

where $\hat{\boldsymbol{\delta}} = \sqrt{T}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Phi}})^t$. The model with the smallest value of $\mathrm{FIC}_{S,p}$ is selected. This version of FIC can simultaneously select a subset of the explicative variables $x_{t,1}, \ldots, x_{t,k}$ and the autoregressive order p of the error term, where $0 \leq p \leq P$.

We illustrate this approach by revisiting the US liquor sales example used in Section 3.5. Now we do not start working with the stationary series, but use the following 'maximal' model for the logarithmic transform of the US liquor sales series:

$$Z_t = \alpha + \beta t + \gamma_2 S_2 + \cdots + \gamma_{12} S_{12} + u_t,$$

where S_i are monthly dummy variables (January is the reference category), and

$$u_t = \phi_1 u_{t-1} + \cdots + \phi_{p_T} u_{t-p_T} + \varepsilon_t.$$

The regression variables which we consider are the constant term, a time trend t, and the set of monthly dummy variables jointly. All information criteria (FIC, AIC, and BIC) agree that the regression variables must all be included. Note that the FIC has now been computed according to (3.13). We again observe that the prediction performance of the criteria is about the same as the simpler situation treated earlier in Section 3.5.

We can extend the idea of simultaneously selecting regression variables and auto-regressive order of the residual series even further. For example, we can allow the variance structure of the residuals to change over time as in the GARCH model by Engle (1982) and Bollerslev (1986). Going one step beyond that, we can include lagged versions of the exogenous variables in the model, such as in the ARX-GARCH model proposed by So et al. (2006). The main ingredient of the FIC in this model is the information matrix $\boldsymbol{J}_{T,\mathrm{full}}$ of the largest model under consideration. (See Claeskens and Hjort, 2003, Sections 2 and 3 for more details). An unbiased empirical estimate thereof can be obtained by using the negative Hessian of the log-likelihood, evaluated at its minimum.

3.7 Conclusions

In this paper we extended the FIC mechanism to allow for an increasing number of parameters as the sample size increases. We specifically worked inside the

framework of h-step ahead prediction of time series using an AR-model, with the direct or plug-in methods for prediction. We illustrated, via simulations, that FIC selects models which give predictions with a comparable MSE to that of AIC over the entire parameter space. This observation holds for both the single-series and the two-series case. This simulation study also demonstrated that the relative mean squared errors for the plug-in method for prediction are quite comparable to those of the direct method. We gave a theoretical justification for Hansen's (2005) use of the FIC for the impulse response, and illustrated that FIC gives better estimates for the impulse response function in certain areas of the parameter space. An extension to simultaneous selection of regression variables and autoregressive order is promising for exploring in greater depth.

Chapter 4

An Information Criterion for Variable Selection in Support Vector Machines

This article has been submitted as

Claeskens, G., Croux, C., and Van Kerckhoven, J. (2007). An Information Criterion for Variable Selection in Support Vector Machines.

Abstract

Support vector machines for classification have the advantage that the curse of dimensionality is circumvented. It has been shown that a reduction of the dimension of the input space leads to even better results. For this purpose, we propose two information criteria which can be computed directly from the definition of the support vector machine. We assess the predictive performance of the models selected by our new criteria and compare them to existing variable selection techniques in a simulation study. The simulation results show that the new criteria are competitive in terms of generalization error rate while being much easier to compute. We arrive at the same findings for comparison on some real-world benchmark datasets.

4.1 Introduction

We study classification using the support vector machine (SVM). We start from a training set $\{(x_i, y_i)\}$ containing n observations. Each p-dimensional observation $x_i = (x_{i1}, \ldots, x_{ip})$ has a class label y_i assigned to it, which is either $+1$ or -1. We wish to find a function $f(\cdot)$ such that for an observation x the predicted class $\hat{y} = +1$ if $f(x)$ is positive, and $\hat{y} = -1$ if $f(x)$ is negative. We want this function to assign the correct class labels to the training observations (low training error rate) and to accurately classify new observations (low generalization error rate). Working with a subset of the p variables $x_{i1}, \ldots, x_{ip}$ reduces variability of the class-label estimator and might lead to better out-of-sample predictions.

It is only true to some extent that variable selection would not be necessary in the support vector machine setting since it manages to circumvent the so-called "curse of dimensionality" (see for example Cristianini and Shawe-Taylor, 2000, Hastie, Tibshirani, and Friedman, 2001, or Schölkopf and Smola, 2002). While the SVM approach avoids fitting a number of parameters equal to the dimension of the input space, there remains the high probability of a perfect separation in high-dimensional problems. For example, if p is larger than the number of observations, it is always possible to perfectly separate the two classes of training data by a hyperplane. In general, the risk of overfitting will increases with the dimension for most data configurations. Hence, the risk of obtaining a decision rule with poor generalization properties (high generalization error rate) cannot be avoided. Guyon et al. (2002) illustrate this and show that variable selection can further improve the SVM's performance.

Variable selection techniques can be divided into three categories. Filters select subsets of variables as a pre-processing step, independently of the prediction method. Wrappers utilize the classification method to score subsets of variables. Finally, embedded methods include variable selection into the construction of the classifier. In this paper we propose new information criteria for SVMs, yielding a wrapper method where we consider the SVM merely as a black box. We refer to Guyon and Elisseeff (2003) for an introduction to variable and feature selection in Machine Learning. Information criteria are a standard tool for model selection in

traditional statistics. Information criteria for variable selection assign a numerical value to each subset of the variables under consideration. The subset with the lowest value of the information criterion is then selected. Examples are the Akaike information criterion (AIC, Akaike, 1973) and the Bayesian information criterion (BIC, Schwarz, 1978). Claeskens and Hjort (2008) survey and explain the use of common information criteria for statistical variable selection in likelihood-based models, we refer to there for more references.

For support vector machines only very few information criteria have been developed. The kernel regularisation information criterion (KRIC) of Kobayashi and Komaki (2006) was originally proposed for parameter tuning of the SVM. We apply it for variable selection. However, the KRIC has a complicated definition and is computationally expensive for large sample sizes. In this paper two new information criteria are proposed, one shares properties with AIC, the other with BIC. We want the new criteria to select a preferably compact subset of variables with good predictive properties. We will show that submodels selected by the new criteria are as performing as the ones chosen by the KRIC, while they incur substantially less computational overhead. We also make a comparison with using cross-validated error rate based criteria, as in Kearns et al. (1997). An important contribution of this paper is that our numerical comparisons show that the popular, but time consuming, cross-validation criteria are outperformed in generalization error by the new information criteria, where the latter are coming at almost no additional computational cost.

Alternative approaches perform variable selection in feature space instead of in input space (Shih and Cheng, 2005), or select a set of "maximally separating directions" in the input space Fortuna and Capson (2004). These methods, however, do not select a set of original input variables. Various other authors have suggested different formulations for the SVM such that variable selection is performed automatically. Examples of such embedded methods can be found in Bi et al. (2003), Zhu et al. (2004), Neumann, Schnörr and Steidl (2005), Lee et al. (2006), Wang, Zhu, and Zou (2006), Zhang (2006), and Lin and Zhang (2006).

In Section 4.2 we define the support vector machine setting, we review ex-

isting information criteria and we describe ranking techniques to speed up the variable selection process. In Section 4.3, we define the new information criteria and highlight their advantages. Section 4.4 contains the results of a simulation study and in Section 4.5 we compare the different techniques on a few real-world benchmark datasets. Section 4.6 concludes and gives some directions for further research.

4.2 Problem setting

4.2.1 The support vector machine

We denote the training sample (x_i, y_i), $1 \leq i \leq n$, with x_i a p-dimensional vector of explicative variables, and $y_i \in \{-1, +1\}$ the class label. The goal is to estimate a target function $f(x)$ in the space of explicative variables such that $f(x_i) > 0$ for $y_i = +1$, and $f(x_i) < 0$ for $y_i = -1$.

We start with linear support vector machines, where $f(x)$ is of the form $f(x) = w'x + b$. For binary classification this function is obtained by solving the minimisation problem

$$\min_{w,b,\xi_i} \left\{ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i \right\} \text{ subject to } \begin{cases} y_i(w'x_i + b) \geq 1 - \xi_i, \\ \xi_i \geq 0, i = 1, \ldots, n. \end{cases} \tag{4.1}$$

The ξ_i are slack margin variables, indicating how close a point x_i lies to the separating boundary (if $\xi_i < 1$), or how badly it is misclassified (if $\xi_i > 1$). The tuning parameter C controls how much weight is put on trying to achieve perfect separation.

The dual problem can be solved more easily, and has the following form:

$$\min_{\alpha}\{\frac{1}{2}\alpha' Q\alpha - \sum_{i=1}^{n} \alpha_i\} \text{ subject to } \begin{cases} 0 \leq \alpha_i \leq C, \quad i = 1, \ldots, n, \\ \sum_{i=1}^{n} y_i\alpha_i = 0. \end{cases} \tag{4.2}$$

Here α_i is the weight given to the observation (x_i, y_i), and Q is a positive semi-definite matrix with entries $Q_{i,j} = y_i y_j x_i' x_j$. The vector w can be found from

$w = \sum_{i=1}^{n} y_i \alpha_i x_i$. The negative intercept b is found by computing $b = 0.5(r_2 - r_1)$, where

$$r_1 = \frac{\sum_{0<y_i\alpha_i<C}(Q\alpha)_i - 1}{\sum_{0<y_i\alpha_i<C} 1} \text{ and } r_2 = \frac{\sum_{0>y_i\alpha_i>-C}(Q\alpha)_i - 1}{\sum_{0>y_i\alpha_i>-C} 1}.$$

If no i exist for which $0 < y_i\alpha_i < C$, then define

$$r_1 = \frac{1}{2}\Big(\min_{\alpha_i=0, y_i=1}(Q\alpha)_i - \max_{\alpha_i=C, y_i=1}(Q\alpha)_i \Big),$$

and analogously for r_2, with $y_i = -1$. Note that we can write $\xi_i = [1 - y_i a_i]_+$, where $[x]_+ = \max\{0, x\}$ and where $a_i = f(x_i)$.

The linear SVM can be extended towards more complex decision functions in a rather straightforward way. Therefore we replace the inner products $x_i' x_j$ in the definition of Q by a more general kernel function $K(x_i, x_j)$. See Cristianini and Shawe-Taylor (2000) for the properties that these kernel functions must have. This leads to a more general decision function

$$f(x) = \sum_{i=1}^{n} y_i \alpha_i K(x_i, x) + b. \tag{4.3}$$

Popular choices for the kernel function in (4.3) are the linear kernel, where the kernel function is $K(x, z) = x'z$, the polynomial kernel of the form $K(x, z) = (c_0 + \gamma x'z)^d$, and the radial basis kernel $K(x, z) = \exp(-\gamma\|x - z\|^2)$, where c_0, γ and d are regularization parameters that can be tuned for optimal performance of the classifier. In this more general setting, we have

$$\|w\|^2 = \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j K(x_i, x_j) = \alpha' Q \alpha$$

for the squared norm of the weight vector, where $Q_{i,j} = y_i y_j K(x_i, x_j)$.

4.2.2 Existing variable selection techniques

We compare our new methods (Section 4.3) to variable selection based on (ten-fold) cross-validation (CV), guaranteed risk minimisation (GRM, Vapnik 1982)

and the kernel regularisation information criterion (KRIC) by Kobayashi and Komaki (2006). Each of these will be explained in more detail below.

Ten-fold cross-validation divides the training data in ten parts of roughly equal size. One part is left out, the other nine parts are the training data and are used to fit the SVM. This SVM is applied to the part that is left out to obtain an estimate of the error rate. This process is repeated ten times (each time a different part is left out) to obtain the CV generalization error rate $\hat{\varepsilon}(S)$ as the average of the ten separate error rates. We select the model with the lowest value of $\hat{\varepsilon}(S)$, where S ranges over all subsets of variables under consideration. Another common method is five-fold CV. The lower the number of folds, the less computing time is required, but the higher the variability of the estimates of the generalization error. Note that $n-$fold CV is the same as the computationally infeasible leave-one-out CV.

General risk minimisation (Vapnik, 1982) is derived from the estimated generalization error rate, using

$$GRM(S) = \hat{\varepsilon}(S) + \frac{|S|}{n}\big(1 + \sqrt{1 + \hat{\varepsilon}(S)(n/|S|)}\big). \tag{4.4}$$

Here, $|S|$ stands for the number of input variables in the set S and n is the number of observations in the training sample. We select the model with the lowest value of $GRM(S)$, where S ranges over all subsets of variables under consideration. Kearns et al. (1997) compare CV, GRM and minimum description length (Rissanen, 1989). Their experiments have demonstrated that none of the criteria is consistently better than the others. Note that the computational overhead for computing these measures can be immense, since we need to train ten support vector machines to estimate the generalization error rate for only one submodel.

We now define the KRIC of Kobayashi and Komaki (2006). This criterion was originally developed to tune the constant C in the SVM definition (4.1), and by extension to tune the kernel parameters. We use it without much adjustment for variable selection. Denote by $x_{i,S}$ the subvector of x_i, consisting of elements x_{ij} with $j \in S$, and similarly for other vectors. We estimate the SVM (4.1) using the observations $(x_{i,S}, y_i)$, yielding the vectors ω_S, b_S and ξ_S, where the subscript

S refers to the subset of variables under consideration. In the dual problem (4.2), we have $\alpha_S = (\alpha_{S,1}, \ldots, \alpha_{S,n})$ and $[Q_S]_{i,k} = y_i y_k K(x_{i,S}, x_{k,S})$. The decision rule $f_S(x)$ is as in (4.3), and we set $a_{i,S} = f_S(x_{i,S})$. Next, we define vectors t_S and m_S of length n, with components

$$t_{S,i} = \eta^2 \frac{\exp(-\eta a_{i,S} y_i)}{(1+\exp(-\eta a_{i,S} y_i))^2} \quad \text{and} \quad m_{S,i} = -\eta \frac{y_i \exp(-\eta a_{i,S} y_i)}{1+\exp(-\eta a_{i,S} y_i)}, \;\; i = 1, \ldots, n.$$

Here we choose $\eta = \log(2)$ such that $\log(1+\exp(-\eta x))$ and $\eta[1-x]_+$ coincide for $x = 0$, see Kobayashi and Komaki (2006) for further motivation. With $\lambda = C^{-1}\log 2$ the KRIC for the logistic Bayesian model for SVMs is defined as

$$\begin{aligned} \mathrm{KRIC}(S) \;\; = \;\; & 2\Bigg[\sum_{i=1}^{n} \log\left(1+\exp(-\eta a_{i,S} y_i)\right) \\ & + \mathrm{trace}((Q_S \mathrm{diag}(t_S) + \lambda I_n)^{-1} Q_S (\mathrm{diag}(m_S)^2 - n^{-1} m_S m_S^t)) \Bigg]. \end{aligned} \tag{4.5}$$

Alternatively, Sollich's Bayesian model for SVMs (Sollich, 2002) leads to a KRIC with a similar form as the one in (4.5). Using

$$\nu(a_{i,S}) = (1+\exp(-2C))^{-1}(\exp(-C[1-a_{i,S}]_+) + \exp(-C[1+a_{i,S}]_+)),$$

the KRIC for the Sollich Bayesian model for SVMs is defined as

$$\mathrm{KRICS}(S) = \mathrm{KRIC}(S) - 2n \log \sum_{i=1}^{n} \nu(a_{i,S}). \tag{4.6}$$

The computation of the KRIC includes inverting an $n \times n$-matrix with only a few zeroes. Therefore, the computation is time-consuming if the sample size n is large. Both the CV error rate and the KRIC may require a prohibitive computing time when a large number of different models needs to be evaluated.

4.2.3 Ranking techniques

A full subset search is computationally not feasible even not for problems with only a small number of dimensions ($p = 15$ for example). To dramatically reduce the number of models while still selecting a model that is "almost" the best model,

Chen, Li and Li (2005) use a genetic algorithm, while Peng, Long and Ding (2005) suggest a combined backward elimination/forward selection strategy. However, both of these techniques still suffer from the possibility that a large number of models needs to be checked before arriving at a solution.

Alternatively, variable ranking consists of assigning a "value of importance" to each variable and sorting the variables according to their importance. This results in a series of p stacked models, thus only p evaluations of the variable selection criterion are needed. The most commonly used algorithm is the SVM recursive feature elimination (SVM-RFE) technique from Guyon et al. (2002). For a linear SVM, the variables are ranked by w_j^2, with w_j the j-th component of the weight vector w. This technique assumes that the variables are standardized to have mean 0 and variance 1. The extension proposed by Rakotomamonjy (2003) allows application to SVMs with a non-linear kernel. We use the following SVM-RFE algorithm with variable influence

$$\Delta\|w_S\|^2_{(j)} = \left|\|w_S\|^2 - \|w_{S\setminus\{j\}}\|^2\right|$$

as suggested by Rakotomamonjy (2003).

Step 1: Initialise $S \leftarrow \{1, \ldots, p\}$, the subset of unranked features, and $r \leftarrow ()$, the vector of ranked features.

Step 2: Repeat the following steps until $S = \emptyset$.

(a) Train a SVM on $(x_{i,S}, y_i)$, and compute $\|w_S\|^2 = \alpha'_S Q_S \alpha_S$.

(b) For each $j \in S$, train a new SVM on $(x_{i,S\setminus\{j\}}, y_i)$. This gives a value $\|w_{S\setminus\{j\}}\|^2 = \alpha'_{S\setminus\{j\}} Q_{S\setminus\{j\}} \alpha_{S\setminus\{j\}}$ for each $j \in S$.

(c) Obtain $j_0 = \text{argmin}_j \left|\|w_S\|^2 - \|w_{S\setminus(j)}\|^2\right|$ and set $S \leftarrow S \setminus \{j_0\}$ and $r \leftarrow (j_0, r)$.

The vector r contains the ranked variables, with the first element the most important one. A disadvantage of this method is that the number of SVMs to be trained is $\mathcal{O}(p^2)$. This can be overcome by using α_S instead of $\alpha_{S\setminus\{j\}}$ in Step 2b, such that $\|w_{S\setminus\{j\}}\|^2 \approx \alpha'_S Q_{S\setminus\{j\}} \alpha_S$. Rakotomamonjy (2003) argues that this

will not affect the ranking significantly, while still allowing a major reduction in computational time, bringing the number of SVMs to be estimated to $\mathcal{O}(p)$. We employ this approximation in the simulation study in Section 4.4 and in the real data examples in Section 4.5.

The most easiest way to rank the variables is by filtering methods. Zhang et al. (2006) propose using $s_j = |w_j(m_{j,+1} - m_{j,-1})|$ for ranking, where $m_{j,+1}$ and $m_{j,-1}$ are the within-class means of variable j. Shih and Cheng (2005) use the Fisher score

$$S_j = \frac{|m_{j,+1} - m_{j,-1}|}{\sqrt{\sigma^2_{j,+1} + \sigma^2_{j,-1}}}$$

for a linear SVM, where $\sigma^2_{j,+1}$ and $\sigma^2_{j,-1}$ are the within-class variances of variable j. The main advantage of using S_j is that it is not necessary to train any SVM to rank the variables. The Fisher score ranking is considered in Sections 4.4 and 4.5.

4.3 The new information criteria

As stated in the previous section, evaluating the CV error rate or the KRIC of a particular support vector machine model requires a high number of additional computations. For this reason, we propose two new criteria which use information already available in the SVM, without additional complicated computations. The criteria are based on how badly the SVM violates the margin constraints, which are written as $\sum_{i=1}^{n} \xi_{i,S}$, where $\xi_{i,S}$ is the margin slack of observation i in the support vector machine trained on the variables with indices in S, where S is a subset of $\{1, \ldots, p\}$. Alternatively, we can use the logarithm of this sum, analogous to Bai and Ng (2002) for selecting the number of factors in factor analysis. However, in the SVM setting this has the drawback that the value is undefined if the sum equals zero, which can happen if the data are perfectly separable. Also, Bai and Ng (2002) advise using a log-transform for scalar invariance reasons. Since we follow the advice to standardise the variables before training the SVM, for better ranking as explained in Section 4.2.3, we automatically have scalar

invariance of the sum of the margin slacks. For these reasons, we choose not to take the log-transform.

Generally (but not always), $\sum_i \xi_{i,S}$ will decrease as more variables are added. Therefore we add a penalty term related to the number of included variables to ensure a tradeoff between accuracy and simplicity of the chosen model. We suggest adding a linear penalty term, such that we get an information criterion of the form

$$IC(S) = \sum_{i=1}^{n} \xi_i + C(n)|S|\,, \tag{4.7}$$

where S is the set of variables included in the model.

A first choice is to take $C(n)$ constant in (4.7). It is interesting to note that $IC(S)$ is then, up to constant factors, an easily computable approximation of the KRIC of Kobayashi and Komaki (2006), hereby providing a theoretical justification for its use. To better understand this, note first that $\log\left(1+\exp(-\eta a_{i,S} y_i)\right)$ is a continuous approximation of the hinge loss function $\eta[1-y_i a_{i,S}]_+ = \eta\xi_{i,S}$ for all $1 \leq i \leq n$. Hence, the first term of the KRIC can be approximated, up to a constant factor, by $\sum_i \xi_{i,S}$. For the approximation of the second term in (4.5), rewrite

$$\begin{aligned} W &= (Q_S \mathrm{diag}(t_S) + \lambda I_n)^{-1} Q_S (\mathrm{diag}(m_S)^2 - n^{-1} m_S m_S^t) \\ &= V \mathrm{diag}(t_S)^{-1} (\mathrm{diag}(m_S)^2 - n^{-1} m_S m_S^t), \end{aligned}$$

with $V = (A + \lambda I_n)^{-1} A$ a symmetric, positive semi-definite matrix and $A = Q_S \mathrm{diag}(t_S)$. Denoting A^- the generalised inverse of A, and using a series expansion around $\lambda = 0$, gives that the leading term of $V = A^-(I + \lambda A^-)^{-1} A$ is equal to $A^- A$. This expansion converges as long as the eigenvalues of λA^- are strictly less than one, which can be obtained by taking λ small enough. We now use a singular value decomposition of both A and A^- and use the fact that the singular values of A^- are the reciprocals of the non-zero singular values of A, to obtain that the product $A^- A$ is a $n \times n$ diagonal matrix with on the diagonal $|S|$ ones and the remaining entries zero. Thus, the leading term of trace(W) equals the sum of $|S|$ diagonal entries of the matrix $\mathrm{diag}(t_S)^{-1}(\mathrm{diag}(m_S)^2 - n^{-1} m_S m_S^t))$.

The i-th diagonal element of this matrix is equal to

$$\frac{n-1}{n} t_{S,i}^{-1} m_{S,i}^2 = \frac{n-1}{n} \exp(-\eta a_{i,S} y_i).$$

To further facilitate computations we replace this by 1, motivated by the fact that $\eta a_{i,S} y_i$ is often small. Although this approximation might be crude for a single term, we found empirically that it works well for the summation over the entire training set. Hence, we arrive at the approximation $\text{trace}(W) \approx |S|$ which is the linear penalty term in (4.7).

Taking the constant value $C(n) = 2$, leads to our first new support vector machine information criterion (SVMIC):

$$\text{SVMICa}(S) = \sum_{i=1}^{n} \xi_i + 2|S|. \tag{4.8}$$

The newly proposed criterion SVMICa for support vector machines shares the form of the penalty with the well-known Akaike (1973) information criterion. This AIC is defined as minus twice the value of the maximised log likelihood of the model, plus two times the number of parameters to be estimated (that is, $2|S|$). Because the penalty $2|S|$ is not dependent on the sample size n, we expect that both criteria share some properties, such as having the tendency to not select the most parsimonious model. For the AIC, Woodroofe (1982) has shown that in the limit for $n \to \infty$, the expected number of superfluous parameters is less than one.

To support the definition of SVMICa , we ran a simulation experiment and compared the values of KRIC and SVMICa for 100 models. The sample size is $n = 50$, with 10 variables of which only the first 4 variables are different from zero. A detailed description of the simulation setting can be found in Section 4.4. We used a linear kernel. Figure 4.1 reports these numerical results and shows a high correlation (0.975) between the values of the two criteria. Other simulation settings gave comparable correlation values.

Our second proposed criterion follows the spirit of Schwarz's (1978) Bayesian information criterion (BIC). This criterion is defined similarly as the AIC, but instead of the penalty $2|S|$, it uses $\log(n)|S|$. The BIC has been shown to be

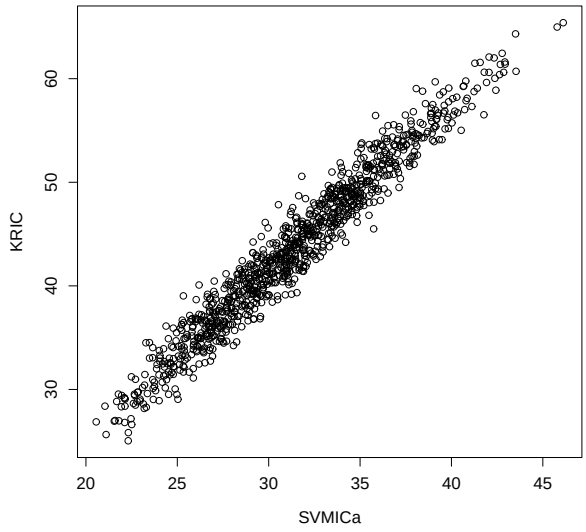

Figure 4.1: Values of KRIC and SVMICa in a simulation experiment, showing high correlation (0.975).

consistent (Haughton 1988, 1989). This means that if the true model is contained in the search list, the criterion will (in the limit for $n \to \infty$) select this correct model. For a related construction for factor models, see Bai and Ng (2002). This motivates us to take $C(n) = \log(n)$, and we define our second criterion

$$\text{SVMICb}(S) = \sum_{i=1}^{n} \xi_i + \log(n)|S|. \tag{4.9}$$

It is immediate that the computational cost of both SVMICs is much lower than of the cross-validated error rate (10 more SVMs to train for 10-fold cross-validation) and of the kernel regularisation information criterion KRIC (which needs computations of the order $\mathcal{O}(n^3)$ due to the matrix inversion). The best case is when the $\xi_{i,S}$ are directly available. Computing the SVMICs is only an $\mathcal{O}(n)$ computation in that case, and usually even less when employing the property that

$$\xi_{i,S} \neq 0 \Leftrightarrow \alpha_{i,S} = 1.$$

When only α_S and Q_S are available, $\xi_{i,S}$ is computed using the relation

$$\xi_{i,S} = \Big[1 - y_i \sum_{\substack{j=1 \\ \alpha_{j,S}>0}}^{n} \alpha_{j,S}[Q_S]_{ij}\Big]_+.$$

This means that in the worst case, the computation time of the $SVMIC$s is $\mathcal{O}(n^2)$, which is still faster than using either CV error rate or KRIC.

4.4 Simulation results

We perform $M = 100$ simulation runs with the following settings. We generate $n \in \{25, 50, 100, 200\}$ independent observations x_i, $1 \leq i \leq n$ of dimension $p \in \{25, 50, 100, 200\}$, with distribution $\mathcal{N}(0, \sigma^2 I_p)$ where $\sigma^2 = 1$. For each observation we generate a class label $y_i \in \{-1, +1\}$, with $P(y_i = 1) = 1/2$. Finally, we let $\mu = (1/2, -1/2, -1/2, 1/2, 0, \ldots, 0)$ of dimension p, and set $x_i \leftarrow x_i + y_i\mu$ to separate the two classes to some extent. This implies that the optimal separating hyperplane is $x'\mu = 0$, such that $\hat{y} = +1$ if $x'\mu > 0$, resulting in a generalization error rate of $\Phi(-\|\mu\|_2/\sigma)$, with Φ the cumulative distribution function of a standard normal. In our example, with $\sigma = 1$ and $\|\mu\|_2 = 1$, we find an optimal generalization error rate of 0.159.

During each simulation run, we standardize the variables to improve the numerical performance of the SVM algorithm. The variables are ranked using either the Fisher score or based on the variable influence on w, as described in Section 4.2.3. For each of the nested models obtained in the variable ranking step, we compute (i) SVMICa and (ii) SVMICb as in (4.8) and (4.9). We compare their performance to (iii) ten-fold CV, (iv) Vapnik's GRM as in (4.4), (v) KRIC for the logistic Bayesian model for SVMs as in (4.5), and (vi) KRIC for the Sollich model for SVMs as in (4.6). An important remark is that for ten-fold CV, we employ the CV2 method, which includes the feature selection procedure in each cross-validation step, as suggested by Zhang et al. (2006). Computing the CV error rate in the usual way can lead to a (severely) biased estimate of the generalization error, and using CV2 reduces this bias.

The experiment is repeated with two different kernels (i) a linear kernel $K(x_1, x_2) = x_1' x_2$ leading to a linear decision rule (ii) a quadratic kernel $K(x_1, x_2) = (\gamma x_1' x_2 + 1)^2$, with $\gamma = 1/p$, the inverse of the number of variables, leading to a quadratic decision rule. The tuning parameter C in each SVM that we train is chosen to be $C = 1$, as we standardize the explicative variables a priori. This is also the standard setting for C for the `svm` procedure in the `R` software package. We experimented with other values of C in the range from 0.1 up to 10, and found only minor differences in the simulation outcomes. We test the accuracy of the classifiers computed from the selected input variables by estimating their generalization (out-of-sample) error rate from a test sample of 10000 new observations. These observations are generated in the same way as the training sample.

Table 4.1 reports the generalization error rates, obtained by averaging over the 100 simulation runs. An overall observation is that the error-rate based selection criteria (CV and GRM) have the worst performance. The performances of the KRICs and the new SVMICs are comparable. More precisely, we observe that the KRICs are better as a variable selection method for small sample sizes ($n = 25$), while the SVMICs give better results for larger sample sizes. This is especially apparent when the quadratic kernel is used. For a small number of observations compared to the number of variables, we also note that SVMICa slightly outperforms SVMICb in terms of generalization error rate, and that the opposite is true with many observations and fewer variables. The differences in generalization error rates become smaller as the number of variables grows. This is particulary true for CV, whose relative performance becomes better at large sample sizes. But SVMICa and SVMICb are still somewhat ahead, and have the advantage that they are much easier (and less time-intensive) to compute than the other criteria, included the KRICs having a computation time of order $\mathcal{O}(n^3)$. Note that, as n grows, the generalization error rates of the models obtained by our two suggested criteria are converging towards the theoretically obtained minimal generalization error rate of 15.9%. Investigating which variable ranking criterion is better, results in case of linear kernels to a strong preference for ranking with the Fisher score. For the quadratic kernel, it is slightly better to rank the variables based on variable influence on $\|w\|^2$.

Linear kernel													
n	p	SVMICa		SVMICb		CV		GRM		KRIC		KRICS	
25	25	32.2	29.4	32.6	31.6	33.5	31.8	36.2	34.5	31.3	29.0	31.5	29.9
	50	34.6	31.6	35.3	32.6	35.3	33.5	37.4	35.4	34.4	33.2	34.4	33.2
	100	37.4	33.9	37.3	35.0	37.8	34.4	38.6	35.7	37.0	34.9	37.1	34.9
50	25	24.4	21.6	24.6	23.2	27.1	25.5	31.1	29.6	25.7	24.9	26.0	25.9
	50	28.5	23.3	27.7	24.8	29.5	26.3	31.4	30.5	29.8	28.7	30.2	29.7
	100	30.9	24.6	29.1	25.0	31.0	28.0	32.1	30.9	31.0	30.1	31.3	30.8
100	25	19.9	18.5	19.6	18.9	24.6	23.8	30.1	30.1	21.8	20.6	22.3	21.7
	50	22.9	19.2	20.2	19.0	25.8	25.4	29.9	29.6	26.9	26.8	27.3	27.8
200	25	17.8	17.0	16.9	16.8	22.7	21.5	28.9	29.3	18.7	18.0	19.2	18.9
Quadratic kernel													
n	p	SVMICa		SVMICb		CV		GRM		KRIC		KRICS	
25	25	31.3	30.7	34.2	33.8	33.8	32.9	37.7	36.6	29.5	28.4	30.2	30.1
	50	35.8	35.3	39.3	38.5	39.6	38.5	43.6	42.6	33.3	33.0	33.9	34.1
	100	43.3	43.3	48.3	48.4	42.8	42.7	49.2	48.7	37.1	37.1	37.7	38.2
50	25	22.7	21.3	25.0	24.3	26.7	25.9	31.8	31.7	23.6	22.5	24.8	25.1
	50	24.4	23.0	26.8	26.8	29.8	28.1	33.9	33.5	27.6	27.1	29.1	29.3
	100	26.4	25.6	30.8	30.2	34.1	33.8	40.3	40.1	31.1	30.9	32.5	32.8
100	25	19.4	18.5	19.9	19.1	23.8	19.2	30.6	30.2	20.0	20.0	21.7	22.0
	50	19.7	18.5	19.8	19.5	24.2	22.0	30.5	30.7	22.6	22.6	24.7	25.1
200	25	20.1	20.3	17.1	16.8	22.4	21.4	29.4	29.6	18.3	18.1	20.3	20.6

Table 4.1: Simulated average generalization error rate (%) for the six methods using two different kernels. For each method, the number on the left resulted from ranking by variable influence on $\|w\|^2$, and the number on the right in each column is from ranking by the Fisher scores S_j.

Figure 4.2 presents the values of the 100 simulated generalization errors as boxplots, giving insight in the variability of the variable selection methods. For most of the cases it turns out that cross-validation is highly variable, while GRM has a small variability. This good property of GRM is, however, accompanied by a much higher average generalization error rate. Comparing the different information criteria shows that SVMICa is quite comparable to the KRICs. The

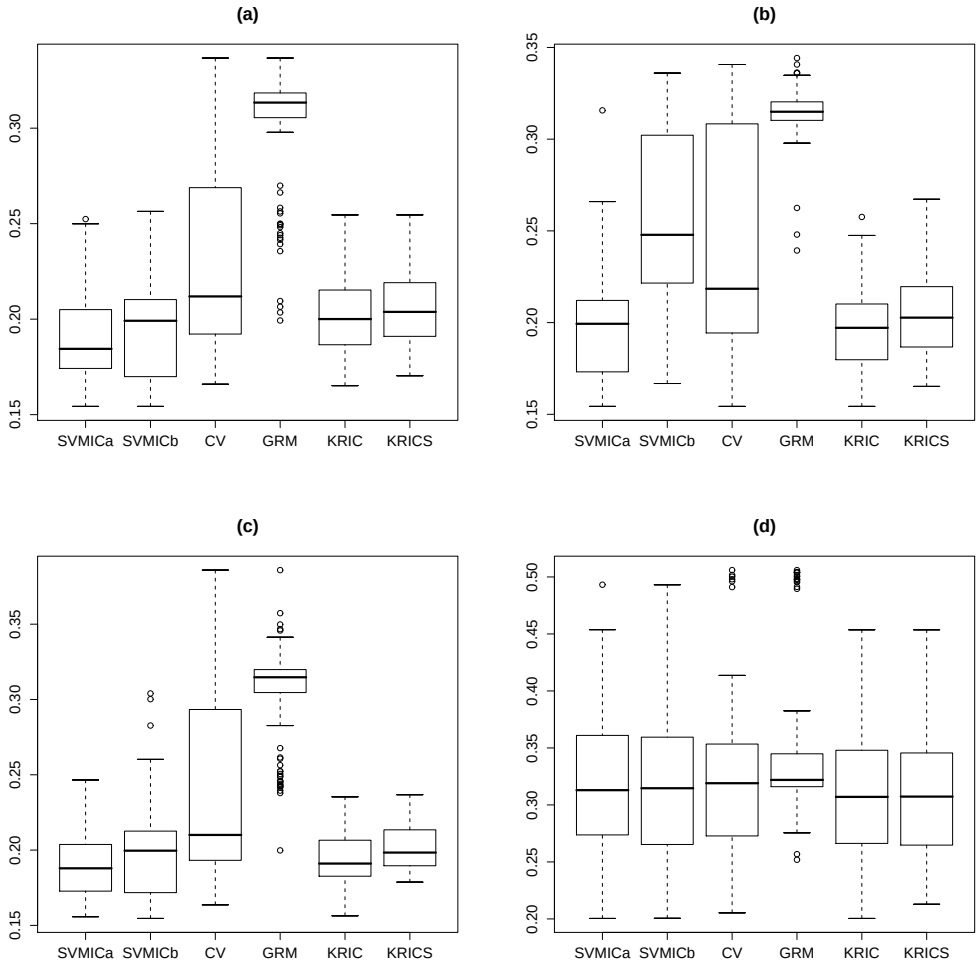

Figure 4.2: Generalization error rates for 100 simulation experiments, for $n = 100$, $p = 25$ (a) linear kernel, ranking with $\|w\|^2$, (b) linear kernel, ranking with Fisher score, (c) quadratic kernel, ranking with $\|w\|^2$, and for (d) $n = 25$, 100 variables, linear kernel and ranking with $\|w\|^2$.

SVMICb has a larger variability. In the setting with small sample size ($n = 25$) and relatively large number of variables (100), all methods, except for GRM, are

comparable with respect to variability, but GRM has again the largest median error rate. Our main conclusion from this analysis is that SVMICa has a similar variability than the KRIC criteria, but SVMICb has a larger variability. Recall that the average error rates, as reported in Table 1, were of similar magnitude for all the four information criteria. Hence, when needing to choosing between the two newly proposed information criteria, we have a preference for SVMICa.

Given the variability of the generalization errors over the 100 simulation runs, see the boxplots in Figure 4.2, it is important to test whether the averages reported in Table 1 are also significantly different from each other. We performed standard t-tests, and most difference are indeed significant. For example, for the settings presented in Figure 1, we obtained that, at the 1% level, (a) all differences are significant, except between SVMICb and the 2 KRiCs (b) all differences are significant, except between SVMICa and the 2 KRICs (c) all differences are significant, except between SVMICb and the 2 KRICs (d) the differences with the GRM method are significant, the others not.

Furthermore, we investigate which models are actually chosen by the different criteria. This information is reported in Table 4.2. For each setting, it shows how many times the correct subset of input variables, containing only the first four input variables, was chosen (C, correct). This table also shows how many times a too-sparse group of variables was selected (U, underfitting), and how many times a too-rich group of variables was chosen (O, overfitting). So an overfit means that all correct variables are selected, but in addition some superfluous ones, while an underfit selects a subset of the important variables, but no irrelevant variables are included. The good performance of SVMICa and SVMICb might be due to the fact that these criteria seem to have the tendency to select a set of variables which includes all significant ones as the number of observations grows. The simulation results indicate that SVMICa behaves like AIC with its tendency to overfit. The SVMICb seems to share the property of BIC that it selects the correct model more often, if at least this true model is one of the possibilities to select from. The cross-validated error rate, and the general risk minimisation in particular, seem to have the tendency to ignore variables which nevertheless are important. As a consequence, the models that these criteria select are of

	Kernel:	Linear				Quadratic			
	Models selected:	C	U	O	R	C	U	O	R
$n = 25$; $p = 25$	SVMICa	1	22	1	76	3	36	0	61
	SVMICb	0	42	0	58	0	64	0	36
	CV	0	38	4	58	1	40	5	54
	GRM	0	77	0	23	0	75	0	25
	KRIC	1	1	7	91	0	1	25	74
	KRICS	0	0	9	91	0	0	49	51
$n = 200$; $p = 25$	SVMICa	22	0	76	2	2	0	98	0
	SVMICb	77	9	10	4	67	14	6	13
	CV	7	48	43	2	4	43	49	4
	GRM	1	98	1	0	1	99	0	0
	KRIC	6	0	93	1	8	0	84	8
	KRICS	1	0	99	0	0	0	100	0
$n = 25$; $p = 100$	SVMICa	0	8	0	92	0	35	0	65
	SVMICb	0	20	0	80	0	63	0	37
	CV	0	23	6	71	0	33	10	57
	GRM	0	56	0	44	0	64	0	36
	KRIC	0	1	0	99	0	0	41	59
	KRICS	0	0	1	99	0	0	56	44

Table 4.2: Simulated frequencies of selected models, with variable ranking done by influence on $\|w\|^2$. Here 'C' denotes correct selection, 'U' is underfitting, 'O' is overfitting, and 'R' for all other situations.

poor predictive quality. The two KRICs of Kobayashi and Komaki (2006) share the overselection property exhibited by SVMICa, but the KRICs select excessive variables even more frequently than SVMICa. This can explain why these criteria perform somewhat worse when the number of observations is large, and why they outperform the proposed SVMICs when the number of observations is small, since the latter tend to underfit the model in the case of few observations.

This concludes the results for the case of two populations coming from an identical distribution, differing only in mean. Another case that we examined is where the variances of the two populations differ from each other. We performed

Linear kernel													
n	p	SVMICa		SVMICb		CV		GRM		KRIC		KRICS	
25	25	28.9	28.0	30.1	29.2	30.4	28.4	32.7	31.6	29.0	27.5	28.8	27.7
	50	33.3	30.2	34.2	31.3	35.1	31.4	35.3	33.1	32.7	30.7	32.5	30.5
	100	35.6	31.5	35.7	32.3	36.0	32.6	36.9	33.7	34.8	32.6	34.8	33.0
	200	36.5	33.2	36.4	34.4	36.4	34.2	36.6	35.6	36.4	33.5	36.1	33.7
50	25	23.3	20.5	23.9	21.9	26.1	24.9	28.9	28.6	24.2	23.6	24.6	24.3
	50	27.1	21.7	25.7	22.7	27.7	25.2	29.1	28.4	27.7	26.8	27.6	27.1
	100	28.3	23.1	27.4	23.7	28.7	25.2	29.9	28.7	28.4	26.7	28.4	27.5
100	25	19.0	17.4	18.1	17.4	22.7	21.5	27.6	27.6	20.5	20.0	21.0	20.9
	50	21.8	17.8	19.3	18.0	23.5	22.7	26.9	27.0	24.8	25.0	25.0	25.5
200	25	17.0	16.1	15.9	15.6	21.4	20.7	27.0	27.0	17.9	17.0	18.3	17.8
Quadratic kernel													
n	p	SVMICa		SVMICb		CV		GRM		KRIC		KRICS	
25	25	29.2	28.9	31.8	31.8	31.8	28.7	35.4	34.7	25.7	24.9	25.8	26.2
	50	35.1	35.8	39.6	40.0	38.1	37.6	42.8	42.4	30.5	30.8	31.3	32.3
	100	42.1	41.7	48.2	48.1	42.2	42.3	49.4	48.7	35.0	36.0	36.2	38.1
	200	50.1	50.1	50.1	50.1	44.7	44.4	50.1	50.1	38.9	40.0	40.4	41.8
50	25	20.5	19.3	23.5	22.2	25.9	24.5	30.6	30.2	19.0	19.1	19.5	19.9
	50	23.1	22.2	26.1	26.2	28.3	27.6	33.2	32.7	23.8	23.9	25.1	26.1
	100	26.5	25.8	30.4	30.4	34.5	33.7	40.5	40.4	28.2	28.8	30.1	32.3
100	25	14.6	15.2	18.5	16.4	20.8	19.9	27.8	27.1	14.2	14.5	14.5	14.9
	50	17.9	17.0	18.4	17.8	22.0	21.5	27.7	28.3	18.1	18.5	19.5	20.3
200	25	9.9	9.8	12.9	13.2	19.6	17.6	29.3	26.8	10.1	10.3	9.7	9.8

Table 4.3: As Table 1, but now for two populations with different variances

a simulation study, in a similar way as the previous one, where the samples have been drawn from $\mathcal{N}(\mu, I_p)$ for class $+1$, and from $\mathcal{N}(-2\mu, 4I_p)$ for class -1.

The results of this simulation are summarized in Tables 4.3 and Table 4.4. We observe similar results as in the case where both populations had equal variance. Selection based on CV error rate and on GRM still perform rather poor. As before, the performances of the KRICs and SVMICs are similar. More precisely, the SVMICs have an improved performance with respect to the KRICs when the sample size is large ($n \geq 50$) and the linear kernel is used, and the KRICs

	Kernel:	Linear				Quadratic			
	Models selected:	C	U	O	R	C	U	O	R
$n = 25$; $p = 25$	SVMICa	0	22	1	77	1	36	0	63
	SVMICb	0	47	0	53	1	57	0	42
	CV	1	40	1	58	1	39	8	52
	GRM	0	76	0	24	0	70	0	30
	KRIC	0	0	6	94	0	0	25	75
	KRICS	0	0	8	92	0	0	50	50
$n = 200$; $p = 25$	SVMICa	11	0	85	4	0	20	0	80
	SVMICb	69	10	16	5	0	45	0	55
	CV	6	56	37	1	0	33	4	63
	GRM	0	100	0	0	0	56	0	44
	KRIC	5	0	93	2	0	0	40	60
	KRICS	0	0	99	1	0	0	53	47
$n = 25$; $p = 200$	SVMICa	0	1	0	99	0	52	0	48
	SVMICb	0	8	0	92	0	54	0	46
	CV	0	22	2	76	0	22	5	73
	GRM	0	46	0	54	0	54	0	46
	KRIC	0	1	0	99	0	0	46	54
	KRICS	0	0	0	100	0	0	56	44

Table 4.4: As Table 2, but now for two populations with different variances

work slightly better for small sample sizes ($n = 25$). For the quadratic kernel, we notice a good performance of the KRICs, which is only matched by SVMICa for larger sample sizes. From Table 4.4 we can again make the same observations as before when the linear kernel is used. For the quadratic kernel the SVMICs have more difficulty selecting all the relevant variables than the KRICs, which explains why the latter criteria have an improved performance here.

We also conducted a simulation experiment where the input variables were strongly correlated. First, the observations were generated as in the first simulation experiment. Then, we applied the transformation

$$x_{ij} = \rho x_{ik_j} + \epsilon_{ij} \text{ with } \epsilon_{ij} \sim \mathcal{N}(0, \rho^2) \text{ i.i.d.}$$

where $i = 1, \ldots, n$, k_j is chosen arbitrarily between 1 and 4, and $4 < j \leq p/2$,

such that about half of the unimportant input variables are correlated with the four important ones. The parameter $|\rho| < 1$ controls the degree of correlation. We have chosen $\rho = 0.8$ and found similar results (not reported) as for the case where the variances of both class-population differ.

4.5 Tests on real data

We compare the performance of the new methods with that of the other discussed criteria on several real-world datasets. We use some of the benchmark datasets used in Rakotomamonjy (2003), and in Rätsch et al. (2001). The datasets used are the Pima Indians Diabetes database (768 observations, 8 variables), the Statlog Cleveland Heart Disease database (303 observations, 14 variables), and Leo Breiman's ringnorm and twonorm datasets (both 7400 observations, 20 variables). These datasets are available from the UCI Machine Learning Repository (the first two), and the Delve Repository (last two). We perform 100 random splits of the data in a training sample and a test sample, where the size of the training sample is chosen as $\sqrt{2n}$, with n the total number of observations in the dataset. We chose the size of the training set such that there is a sufficient amount of observations in the test sample to estimate the generalization (out-of-sample) error rate. The training sample size is relatively small, such that the computation time for the KRIC remains within bounds. For each of these partitions we perform variable selection on the training sample exactly as in the simulation study. We first rank the variables to retain p stacked subsets of input variables, and then use the information criteria to select the variables that best explain the training data. Then, we predict the class labels for the test sample, and use these predictions to estimate the generalization error rate. We use variable ranking based on variable influence on $\|w\|^2$ as well as on Fisher score, and we use a linear, quadratic and radial kernel.

The estimated generalization error rates are presented in Table 4.5 for each dataset and estimation setting. We observe that the KRICs are the preferred choice of variable selection criterion in terms of generalization error rate for the 'twonorm' and 'heart' datasets. For the 'ringnorm' and 'diabetes' datasets the

	Ranking:	Variable influence on $\|w\|$			Fisher scores		
Data	Kernel:	Linear	Quadratic	Radial	Linear	Quadratic	Radial
Diabetes	SVMICa	28.6	28.5	29.2	28.0	28.2	28.4
	SVMICb	29.0	28.9	29.2	28.6	28.5	28.9
	CV	28.6	29.1	29.1	28.8	28.5	29.3
	GRM	29.6	29.7	29.6	29.1	29.2	29.3
	KRIC	28.5	28.2	29.4	27.5	28.1	29.6
	KRICS	28.6	28.5	29.7	28.3	28.6	29.7
Heart	SVMICa	27.0	27.4	27.7	27.6	28.0	28.3
	SVMICb	27.6	28.9	28.9	28.2	29.3	29.5
	CV	27.6	28.6	27.2	26.8	28.0	28.8
	GRM	29.3	30.3	29.4	28.8	30.4	30.6
	KRIC	25.4	23.4	23.8	24.5	23.2	23.8
	KRICS	25.3	23.5	25.2	25.2	23.7	25.0
Ringnorm	SVMICa	31.1	16.4	8.4	30.8	15.6	6.5
	SVMICb	34.9	20.2	13.5	35.2	22.4	13.4
	CV	33.9	32.1	26.6	32.8	25.6	21.2
	GRM	39.2	41.3	38.6	39.3	38.4	37.3
	KRIC	30.1	16.3	6.0	29.6	15.9	4.4
	KRICS	29.9	16.0	3.1	29.2	15.4	2.5
Twonorm	SVMICa	9.9	9.3	11.4	10.1	8.9	9.4
	SVMICb	13.5	14.1	15.9	15.0	15.2	16.0
	CV	20.5	21.0	19.8	21.0	21.1	20.8
	GRM	31.4	31.7	31.6	30.8	31.2	31.3
	KRIC	8.0	7.5	11.0	6.8	6.8	9.2
	KRICS	7.5	6.0	4.0	6.6	5.5	4.8

Table 4.5: Generalization error rates (%) for variable selection applied to four data sets. Two variable ranking schemes and three types of kernel are used for each of the criteria.

difference in performance between the KRICs and our newly proposed SVMICs is less pronounced. The predictive performance of the models selected by SVMICa are for most settings comparable to that of the KRIC, while being much faster to compute. These results are consistent across all settings. The CV error rate

and especially the GRM have a poor performance, which is in line of the results obtained in the simulation.

From these results, and the results obtained in Section 4.4, we suggest to use either the SVMICa or the SVMICb if a preliminary analysis of the data or a priori knowledge indicates that the true decision function is almost linear. When it differs strongly from a linear function, the researcher has a choice between the ease of computation of the support vector machine information criteria, or the somewhat improved predictive performance, though with higher computational cost, of the kernel regularization information criterion.

Finally, we applied the newly proposed information criteria for variable selection to two large data sets, the "Madelon" ($n = 2000, p = 500$) and "Arcene" data ($n = 100, p = 10000$). These data sets were part of the NIPS 2003 feature selection, and are described in detail in Guyon et al (2006). Given the high dimensionality of these data, the variables were ranked according to the Fisher score. We used a linear kernel and computed balanced error rates (BER), that is the average of the error rate of the positive class and the error rate of the negative class. When using SVMICa we obtain a BER of 43.0% for the Madelon data, and 31.1% for the Arcene data. For SVMICb we get 37.3% and 31.1%, respectively. In Guyon et al (2006, 2007) the BER of other feature selection methods is presented, and it turns out that several other methods yield much better performance on these data. A possible explication is that we used a standard SVM, without any optimal tuning of the regularization parameters.

4.6 Conclusions

In this paper we considered the problem of variable selection in support vector machines. We proposed two new information criteria, SVMICa and SVMICb, which allow us to evaluate the suitability of the selected subset of variables for predictive purposes, without much additional computational costs. We provided an argumentation for these criteria, linking SVMICa to the KRIC of Kobayashi and Komaki (2006), and justifying SVMICb with the need for a consistent selection criterion. We demonstrated the effectiveness of these criteria in a sim-

ulation study, where we compared their predictive performance to the KRIC, cross-validation and general risk minimization. Especially for decision functions which are close to an affine function, we found that SVMICa and SVMICb performed the best of all tested criteria, and were also the easiest to compute. For more complicated decision functions, we found that SVMICa still performs well for selecting models with good generalization properties. We repeated the experiment on several real data examples, and the result confirmed the good properties of these newly proposed criteria. In particular we showed that cross-validation criteria are outperformed in generalization error by the new information criteria, where the latter are coming at almost no additional computational cost.

The aim of our paper was to propose an information criterion for a standard SVM. We do not claim that the procedure is outperforming other very advanced feature selection methods, which are not relying on a standard SVM. Obtaining information criteria for other machine learners is an interesting topic for future research. Another research question is how suitable the information criteria are for optimal tuning of the regularization and other parameters of the SVM, without necessarily selecting a subset of input variables. Finally, it would be interesting to continue on the theoretical verification of the good performance of our two proposed criteria, and for example try to obtain consistency results for the SVM information criteria.

Chapter 5

Classification efficiencies of Convex Risk Minimisation methods at the normal model

This article has been submitted as
Claeskens, G., Croux, C., and Van Kerckhoven, J. (2007). Classification efficiencies of Convex Risk Minimisation methods at the normal model.

Abstract

In this paper the asymptotic classification efficiency of a class of binary classification methods known as convex risk minimisation techniques is derived. We computed the classification efficiency of these techniques relative to the well-known classical Fisher Discriminant rule, which is known to be optimal for two normal populations with equal variances. We find that for reasonably balanced classes which are not easily separable, the convex risk minimisation methods have fairly high classification efficiency in this setting.

5.1 Introduction

In this paper we study the classification efficiency of several binary classification techniques belonging to the group of convex risk minimisation methods (CRM; Vapnik, 1998). These methods have the advantage that they are easily applicable to a variety of classification problems, ranging from linear classification to more general classification rules, such as quadratic rules, or Gaussian kernel based classification rules. Other advantages of CRM methods are their ability to deal with high-dimensional problems, and their good generalisation properties. Because of these desirable properties, convex risk minimisation techniques have become popular classification methods. Examples of such methods include (kernel) logistic regression (Wahba, 1999), AdaBoost (Freund and Schapire, 1996; Friedman et al., 2000; Hastie et al., 2001), and support vector machines (Christianini and Shawe-Taylor, 2000). An interesting question that we answer in this paper is which price you pay, in terms of classification efficiency, for this flexibility. We do this by comparing the CRM techniques with Fisher's linear discriminant rule in the setting where Fisher's rule is optimal.

Determining the classification efficiency of a statistical decision rule was proposed by Efron (1975), who compared logistic regression with the Fisher linear discriminant rule. The efficiency was computed for a mixture of two normally distributed populations with equal variances, where Fisher's rule is known to be optimal. Finding a decision rule's classification efficiency is especially important with classifiers robust to outliers, where researchers want to know what price, in terms of efficiency loss, they pay for the robustness of the classifier. Such efficiencies have been computed in Croux, Haesbroeck and Joossens (2008) for the robust logistic discrimination rule, and in Croux, Filzmoser and Joossens (2008) for robust linear discriminant rules.

In this paper we assume the setting of two normal distributions with the same covariance matrix, which allows us to obtain feasible theoretical and analytical results. We use a similar approach to Croux, Filzmoser and Joossens (2008), who computed the asymptotic loss of the classifier using the second order influence function of the error rate of the classifier. Classification efficiencies are

computed for CRM classifiers, with AdaBoost and Support Vector Machines as leading examples. These results constitute the main contribution of the paper, and it turns out that the studied CRM techniques are reasonably efficient with an efficiency above 50% when the population means of the two populations are at a Mahalanobis distance less than 2 of each other, and this for reasonably balanced populations (log-odds ratio < 1 in absolute value).

In Section 5.2 we introduce the notation, Section 5.3 contains the theoretical framework for general risk minimisation problems. We show that the CRM techniques are Fisher-consistent when the class probabilities are equal to 1/2. In Section 5.4 we calculate the asymptotic classification efficiencies for the specific CRM techniques mentioned above. Section 5.5 provides a numerical comparison of the efficiency of the various techniques. Finally, a summary and some conclusions are in Section 5.6. All the proofs in this paper are relegated to the appendix.

5.2 Model Setting

In this section we introduce the model setting and present basic results for the Fisher linear discriminant rule. Let X be a p-variate stochastic variable representing the predictor variables, and let Y be the variable indicating the class label, so $Y \in \{+1, -1\}$. These random variables (X, Y) follow a joint distribution that we denote by H. The observations in the training sample are generated by H.

In this paper, we focus on linear classification rules of the form

$$\hat{Y} = \text{sign}(a + b^t x), \tag{5.1}$$

where a is the intercept, b is the p-dimensional vector of slope parameters, and x is a p-variate observation to classify. We restrict to linear classification rules because we want to benchmark their performance with respect to the classical Fisher's rule, and this in a model setting where the latter is optimal. Since Fisher's rule is linear, we restrict the other classification rules to be linear too.

Moreover, theoretical classification efficiencies are only analytically computable under this linearity assumption.

The performance of the classifier in (5.1) will be measured by means of a *loss-function* $L(\cdot)$. This function is assumed to be positive, continuous, and convex. The *expected risk* of the classification rule is

$$R_{L,H}(a,b) = E_H\big[L\big(Y(a+b^tX)\big)\big]. \tag{5.2}$$

The values of intercept a and slope b that minimise the expected risk are denoted as

$$\big(A_L(H), B_L(H)\big) = \underset{(a,b)}{\operatorname{argmin}}\, R_{L,H}(a,b), \tag{5.3}$$

and the associated discriminant rule is a *convex risk minimisation* rule. This rule is asymptotically equivalent (at the population level) to the finite sample minimisation problem

$$\min_{a,b} n^{-1}\sum_{i=1} L\big(y_i(a+b^tx_i)\big) + n^{-1}\lambda\|b\|_2^2, \tag{5.4}$$

where n is the size of the sample $(x_1,y_1),\ldots,(x_n,y_n)$ which is drawn i.i.d. from the distribution H, and λ is a regularisation parameter. The minimisation problem (5.4) differs from that used by Christmann and Steinwart (2004) for a study of robustness properties and Zhang (2004) for studying consistency. They assumed that the regularisation penalty term does not vanish for growing sample size, by letting λ grow with n.

Each CRM rule depends on both the distribution of the training data, and on the specific loss function. Table 5.1 gives a list of the classification methods used in this paper, with their associated loss functions. Note that only the loss function for least squares is not decreasing, and therefore we will treat it as a separate case in Section 5.4.4.

To investigate the generalisation error, or out-of-sample error rate, we make the model assumption that the data to classify are drawn from H_m, a mixture of 2 normal distributions $H_+ \equiv \mathcal{N}(\mu_+,\Sigma)$ and $H_- \equiv \mathcal{N}(\mu_-,\Sigma)$. The class probabilities $\pi_+ = P_{H_m}(Y=+1)$ and $\pi_- = P_{H_m}(Y=-1)$ are strictly positive. The

Classification method	$L(u)$
AdaBoost	$\exp(-u)$
(Kernel) logistic regression	$\log(1+\exp(-u))$
Support vector machine	$[1-u]_+$
Least squares	$(1-u)^2$

Table 5.1: Commonly used loss functions for general risk minimisation

error rate (ER) of the CRM-rule based on the loss function L computed from training data following the distribution H is then given by

$$\begin{aligned} \mathrm{ER}_L(H) &= \pi_+ P(A_L(H) + B_L(H)^t X < 0 \mid X \sim H_+) \\ &\quad + \pi_- P(A_L(H) + B_L(H)^t X > 0 \mid X \sim H_-). \end{aligned} \tag{5.5}$$

Ideally we have that $H = H_m$, which means that the distribution of the training data is the same as the distribution of the data-to-classify.

At the model distribution (that is $H = H_m$), we know that the Fisher rule is optimal, in the sense of being the classifier with the smallest out-of-sample error rate (Johnson and Wichern, 1998, page 685). The Fisher rule is given by $\hat{Y} = \mathrm{sign}(\alpha + \beta^t x)$, with

$$\begin{aligned} \beta &= \Sigma^{-1}(\mu_+ - \mu_-) \\ \text{and } \alpha &= \log\frac{\pi_+}{\pi_-} - \beta^t \frac{(\mu_+ + \mu_-)}{2}. \end{aligned}$$

The expression for the optimal error rate is then given by

$$\mathrm{ER}_{\mathrm{opt}} = \pi_- \Phi\Big(\frac{\theta}{\Delta} - \frac{\Delta}{2}\Big) + \pi_+ \Phi\Big(-\frac{\theta}{\Delta} - \frac{\Delta}{2}\Big), \tag{5.6}$$

with $\theta = \log(\pi_+/\pi_-)$ the log-odds ratio, and $\Delta^2 = (\mu_+ - \mu_-)^t \Sigma^{-1} (\mu_+ - \mu_-)$ the squared Mahalanobis distance between the two group means.

5.3 General results

In this section we give the general results which are valid for all convex risk minimisation methods with a decreasing loss function $L(\cdot)$. Since all considered

classification models are affine equivariant, we suppose, without loss of generality, that $\mu_+ = -\mu_- = e_1\Delta/2$, with $e_1 = (1, 0, \ldots, 0)^t$, and $\Sigma = I_p$ the $p \times p$ identity matrix. This model is the *canonical model*.

Definition 5.1 *A convex risk minimisation rule, defined by the loss function L, is said to be Fisher consistent at the model distribution if*

$$\mathrm{ER}_L(H_m) = \mathrm{ER}_{\mathrm{opt}}.$$

In section 5.3.1 we obtain sufficient conditions under which CRM classification rules are Fisher consistent. In Section 5.3.2, we provide general expressions of influence functions. Finally, Section 5.3.3 gives the asymptotic loss and the asymptotic relative classification efficiency of the decision rules, computed from the influence functions.

5.3.1 Fisher-consistency of convex risk minimisation methods

Let H_m denote the canonical model distribution, and $A_L(H_m)$ and $B_L(H_m)$ the minimisers of $R_{L,H_m}(a, b)$, defined in (5.2). To prove Fisher consistency, it needs to be shown that

$$A_L(H_m) = C_L(H_m)\theta \text{ and } B_L(H_m) = C_L(H_m)\Delta e_1, \tag{5.7}$$

for a certain scalar constant $C_L(H_m)$. If (5.7) holds, then it readily follows that $\mathrm{ER}_L(H_m) = \mathrm{ER}_{\mathrm{opt}}$, with $\mathrm{ER}_{\mathrm{opt}}$ given in (5.6). The next proposition states the conditions to ensure Fisher consistency for CRM models.

Proposition 5.2 *If $L(\cdot)$ is a positive, continuous, convex, decreasing function, and if the distribution H_m verifies $\pi_+ = \pi_- = 1/2$, then the convex risk minimisation rule defined by the function $L(\cdot)$ is Fisher consistent.*

If the condition $\pi_+ = \pi_- = 1/2$ does not hold, it is not possible to prove Fisher consistency. In that case, we can only prove that the sign of the intercept $A_L(H_m)$ is the same as the sign of θ. Note that this consistency result differs

from the result obtained in Zhang (2004). The difference lies in our assumption that the regularisation penalty term in (5.3) becomes zero at the population level (when n converges to infinity), whereas Zhang (2004) assumes that this term is still significant.

5.3.2 Influence Functions

To study the effect of an (outlying) observation on a statistical functional, such as the error rate, influence functions (Hampel et al, 1986; van der Vaart, 2000, pp. 291–296) are commonly used. The influence function is defined as

$$\mathrm{IF}\big((x,y);\mathrm{ER},H_m\big) = \lim_{\varepsilon\to 0^+} \frac{\mathrm{ER}\big((1-\varepsilon)H_m + \varepsilon\Delta_{(x,y)}\big) - \mathrm{ER}(H_m)}{\varepsilon},$$

where $\Delta_{(x,y)}$ is the Dirac measure putting all its mass in the observation (x,y). The k-th order influence function of a statistical functional T is defined as

$$\mathrm{IF}k\big((x,y);T,H_m\big) = \frac{\partial^k}{\partial\varepsilon^k} T\big((1-\varepsilon)H_m + \varepsilon\Delta_{(x,y)}\big)\Big|_{\varepsilon=0}.$$

For small amounts of contamination in the training data, due to the presence of a possible outlier (x,y), the error rate of the discriminant rule based on $H_\varepsilon = (1-\varepsilon)H_m + \varepsilon\Delta_{(x,y)}$ can be approximated using the Taylor-expansion

$$\mathrm{ER}(H_\varepsilon) = \mathrm{ER}(H_m) + \varepsilon\mathrm{IF}\big((x,y);\mathrm{ER},H_m\big) + \frac{\varepsilon^2}{2}\mathrm{IF}2\big((x,y);\mathrm{ER},H_m\big) + \mathcal{O}(\varepsilon^3).$$

The Fisher discriminant rule is optimal at the model distribution, and we have $\mathrm{ER}_{\mathrm{opt}} = \mathrm{ER}(H_m)$. This also implies that any other discriminant rule, based on a contaminated training sample, can never have an error rate smaller than $\mathrm{ER}_{\mathrm{opt}}$. Hence, negative values for $\mathrm{IF}\big((x,y);\mathrm{ER},H_m\big)$ are excluded. Using the property that $E\big[\mathrm{IF}\big((x,y);\mathrm{ER},H_m\big)\big] = 0$ (Hampel et al, 1986, p. 84), it follows that $\mathrm{IF}\big((x,y);\mathrm{ER},H_m\big) = 0$ almost surely. Hence, the behaviour of the error rate, under small amounts of contamination, is characterised by the second order influence function $\mathrm{IF}2\big((x,y);\mathrm{ER},H_m\big)$, which should be non-negative everywhere.

For decision rules that are optimal under the model distribution H_m, we can use Proposition 2 of Croux, Filzmoser and Joossens (2008) to determine the

second order influence function of the error rate. This proposition states that

$$\begin{aligned}\mathrm{IF2}\big((x,y);\mathrm{ER},H_m\big) &= \frac{\pi_-\Delta}{C_L(H_m)^2}\phi\Big(\frac{\theta}{\Delta}-\frac{\Delta}{2}\Big)\Big(\sum_{k=2}^{p}\Big(\frac{\mathrm{IF}\big((x,y);B_L,H_m\big)^t e_k}{\Delta}\Big)^2 \\ &\quad + \Big(\frac{\mathrm{IF}\big((x,y);A_L,H_m\big)}{\Delta}-\theta e_1^t\frac{\mathrm{IF}\big((x,y);B_L,H_m\big)}{\Delta^2}\Big)^2\Big),\end{aligned}$$

for Fisher consistent decision rules. This implies that we should determine the influence functions of the estimators of the parameters on the various binary classification rules.

Proposition 5.3 *For a Fisher-consistent convex risk minimisation rule, with loss function $L(\cdot)$, the influence functions of the estimators of the parameters can be expressed as*

$$\begin{aligned}\mathrm{IF}\big((x,y);A_L,H_m\big) &= -yL'\big(yC_L(H_m)(\theta+\Delta x_1)\big)\frac{A_2-A_1x_1}{D} \\ e_1^t\mathrm{IF}\big((x,y);B_L,H_m\big) &= -yL'\big(yC_L(H_m)(\theta+\Delta x_1)\big)\frac{A_0x_1-A_1}{D} \\ e_k^t\mathrm{IF}\big((x,y);B_L,H_m\big) &= -yL'\big(yC_L(H_m)(\theta+\Delta x_1)\big)\frac{x_k}{A_0} \qquad (1<k\leq p),\end{aligned}$$

where we define $e_k^t=(0,\ldots,0,1,0,\ldots,0)$ with a one at the kth position, $D=A_0A_2-A_1^2$ and

$$A_j = E_{H_m}\big[L''\big(YC_L(H_m)(\theta+\Delta X_1)\big)X_1^j\big] \text{ for } j=0,1,2. \tag{5.8}$$

Using the expressions obtained in Proposition 5.3, we can prove the following corollary.

Corollary 5.4 *For any convex risk minimisation method defined by a function $L(\cdot)$, the influence functions on the estimators of the parameters, and by extension the second-order influence function on the error rate, is unbounded.*

This result does not contradict the robustness properties obtained in Christmann and Steinwart (2004), who assumed the use of a bounded, continuous kernel. This assumption is violated here as we restrict to the use of a linear kernel.

5.3.3 Asymptotic Relative Classification Efficiencies

Estimating the decision rule at the finite sample level results in a generalisation error rate ER_n. When the training data are from the model H_m, the expected loss in classification performance is given by

$$\mathrm{Loss}_n = E_{H_m}[\mathrm{ER}_n - \mathrm{ER}_{\mathrm{opt}}]$$

where $\mathrm{ER}_{\mathrm{opt}}$ is the error rate of the optimal decision rule, as given in (5.6). Proposition 3 of Croux, Filzmoser, and Joossens (2008) states that at the model distribution H_m, the expected loss in error rate of an estimated optimal discriminant rules satisfies

$$\mathrm{Loss}_n = \frac{1}{2n} E_{H_m}[\mathrm{IF2}((X,Y);\mathrm{ER},H_m)] + o(n^{-1}),$$

where n is the sample size. This leads to the asymptotic loss

$$\text{A-Loss} = \lim_{n\to\infty} n\mathrm{Loss}_n = \frac{1}{2} E_{H_m}[\mathrm{IF2}((X,Y);\mathrm{ER},H_m)]$$

such that

$$\mathrm{ER}_n = \mathrm{ER}_{\mathrm{opt}} + \frac{\text{A-Loss}}{n} + o(n^{-1}).$$

Our goal is now to compare the classification performance of the methods under consideration with the efficient method, Fisher's linear discriminant rule, which corresponds to the maximum likelihood method at the normal model.

We work with the canonical model described in Section 5.3. For a Fisher-consistent classification rule, defined by a loss function L, the asymptotic loss satisfies

$$\begin{aligned} \text{A-Loss}_L(H_m) &= \frac{\pi_-}{2C_L(H_m)^2\Delta}\phi\Big(\frac{\theta}{\Delta}-\frac{\Delta}{2}\Big)\Big(\mathrm{ASV}(A) - \frac{2\theta}{\Delta}\mathrm{ASC}(A,B_1) \\ &\qquad + \frac{\theta^2}{\Delta^2}\mathrm{ASV}(B_1) + (p-1)\mathrm{ASV}(B_2)\Big), \end{aligned} \tag{5.9}$$

where

$$\begin{aligned} \mathrm{ASV}(A) &= E_{H_m}\big[\big(\mathrm{IF}((X,Y);A,H_m)\big)^2\big] \\ \mathrm{ASV}(B_j) &= E_{H_m}\big[\big(e_j^t\mathrm{IF}((X,Y);B,H_m)\big)^2\big], j=1,2 \\ \mathrm{ASC}(A,B_1) &= E_{H_m}\big[\mathrm{IF}((X,Y);A,H_m)e_1^t\mathrm{IF}((X,Y);B,H_m)\big] \end{aligned}$$

depend on the classification rule and on Δ and θ. This follows immediately from the definition of the asymptotic loss. As the number of variables p increases, it is obvious that the asymptotic loss is dominated by the $ASV(B_2)$ term. For a CRM rule with loss function $L(\cdot)$, these asymptotic variances and covariances can be written as

$$\begin{aligned}
\mathrm{ASV}_L(A) &= \frac{1}{D^2} E_{H_m}\big[L'\big(YC_L(H_m)(\theta+\Delta X_1)\big)^2(A_2-A_1X_1)^2\big] \\
\mathrm{ASV}_L(B_1) &= \frac{1}{D^2} E_{H_m}\big[L'\big(YC_L(H_m)(\theta+\Delta X_1)\big)^2(A_0X_1-A_1)^2\big] \\
\mathrm{ASV}_L(B_2) &= \frac{1}{A_0^2} E_{H_m}\big[L'\big(YC_L(H_m)(\theta+\Delta X_1)\big)^2\big] \\
\mathrm{ASC}_L(A,B_1) &= \frac{1}{D^2} E_{H_m}\big[L'\big(YC_L(H_m)(\theta+\Delta X_1)\big)^2(A_2-A_1X_1)(A_0X_1-A_1)\big],
\end{aligned} \tag{5.10}$$

where we used the expressions obtained in Section 5.3.2. It is important to remark that for CRM methods the final expression of the asymptotic loss depends on the first and second derivative on the loss function.

The optimal asymptotic loss obtained by Fisher's linear discriminant rule is

$$\begin{aligned}
\text{A-Loss}_{\mathrm{opt}}(H_m) &= \frac{1}{2\pi_+\Delta}\phi\Big(\frac{\theta}{\Delta}-\frac{\Delta}{2}\Big)\Big\{p+\frac{\Delta^2}{4}+\frac{\theta^2}{\Delta^2}+(\pi_- -\pi_+)\theta \\
&\qquad +(p-1)\Delta^2\pi_+\pi_- +2\theta^2\pi_-\pi_+\Big\}.
\end{aligned} \tag{5.11}$$

This expression has been obtained in Efron (1975). Using (5.11), we can define the asymptotic relative classification efficiency (ARCE) as

$$\mathrm{ARCE}_L(H_m) = \frac{\text{A-Loss}_{\mathrm{opt}}(H_m)}{\text{A-Loss}_L(H_m)}.$$

Due to Fisher's rule being the efficient rule, this ratio cannot exceed one. The closer to one, the more efficient the convex risk minimisation rule is.

5.4 Specific Results

Section 5.3 contains the results for general convex risk minimisation techniques. In this section, we study four specific CRM rules in detail. We start with AdaBoost in Section 5.4.1. We also repeat the analysis of Efron (1975) for logistic

regression in Section 5.4.2 though now using influence functions and arrive the same results. The support vector machine is examined in detail in Section 5.4.3. We finish by investigating least squares in Section 5.4.4.

5.4.1 AdaBoost

AdaBoost (Freund and Schapire, 1996; Friedman et al., 2000; Hastie et al., 2001) is an example of a convex risk minimisation technique for binary classification with loss function $L(u) = \exp(-u)$. This function satisfies all the conditions in Proposition 5.2, thus AdaBoost is Fisher consistent at the normal model with $\theta = 0$. The next proposition states that Fisher consistency holds at any value of θ.

Proposition 5.5 *AdaBoost is Fisher consistent for all θ, and $C_L(H_m) = \frac{1}{2}$ for all choices of $\Delta > 0$ and θ.*

With this result, we may compute the asymptotic loss using (5.9). For AdaBoost, it is possible to derive an analytic expression for the asymptotic loss. After tedious calculations, which can be found in the appendix, we have the result of the following proposition.

Proposition 5.6 *The asymptotic loss for AdaBoost satisfies*

$$A\text{-Loss(Ada)} = \frac{1}{2\Delta\pi_+}\phi\Big(\frac{\theta}{\Delta} - \frac{\Delta}{2}\Big)\exp\Big(\frac{\Delta^2}{4}\Big)\Big(p + \theta + \frac{\theta^2}{\Delta^2}\Big(\frac{\Delta^2}{4} + 1\Big)\Big) \quad (5.12)$$

at the canonical model.

5.4.2 Logistic Regression

For comparative purposes, we analyse the logistic regression setting. Originally, Efron (1975), did not use influence functions for studying the efficiency of logistic regression. Our alternative calculations coincide with those found earlier. Recall that the loss function here is $L(u) = \log\big(1 + \exp(-u)\big)$. Hence, we find that $L'(u) = -\exp(-u)/\big(1 + \exp(-u)\big) = F(u) - 1$, where we used the shorthand

notation $F(u) = \big(1 + \exp(-u)\big)^{-1}$, and that $L''(u) = F(u)\big(1 - F(u)\big)$. For obtaining the quantities A_0, A_1, and A_2 as in Proposition 5.3, we need to compute the expressions

$$\begin{aligned} A_j &= E_{H_m}\big[F(\theta + \Delta X_1)\big(1 - F(\theta + \Delta X_1)\big)X_1^j\big] \\ &= \int_{\mathbb{R}} \frac{\pi_+\pi_-}{\pi_+ \exp(\Delta x_1/2) + \pi_- \exp(-\Delta x_1/2)} \frac{x_1^j}{\sqrt{2\pi}} \exp\Big(-\frac{\Delta^2}{8}\Big) \exp\Big(-\frac{x_1^2}{2}\Big) \mathrm{d}\, x_1, \end{aligned}$$

for $j = 0, 1, 2$. This integral can be evaluated numerically for a given Δ and θ. After tedious calculations (available upon request), we find that the expressions for the asymptotic variances of the parameters in (5.10) reduce to $\mathrm{ASV}(A) = A_2/D$, $\mathrm{ASV}(B_1) = A_0/D$, $\mathrm{ASC}(A, B_1) = -A_1/D$, and $\mathrm{ASV}(B_2) = A_0^{-1}$, with A_j and D as in Proposition 5.3. This confirms the results in Efron (1975).

5.4.3 Support Vector Machine

The loss function for the support vector machine, $L(u) = [1 - u]_+$, satisfies all the requirements in Proposition 5.2. Thus, we know that the SVM is Fisher consistent in the normal model for $\theta = 0$.

However, we have observed empirically that the support vector machine is not Fisher-consistent if $\theta \neq 0$ ($\pi_+ \neq \pi_-$). To illustrate this, we have optimised

$$R_{H_m}(a, b) = E_{H_m}\big[L\big(Y(a + b_1 X_1)\big)\big],$$

with H_m the canonical model for several values of Δ and θ. The results are shown in Table 5.2. This table lists the values of a and $b_1 = b^t e_1$ which optimise $R_{H_m}(a, b)$. We observe that, for $\theta \neq 0$, the values are not in accordance with relation (5.7), and hence, that the support vector machine is not Fisher consistent. The two last lines in the table give the results of the same optimisation, this time when $\theta = 0$. These results verify empirically that the SVM is Fisher consistent for balanced populations. For the remainder of this section, we assume that $\theta = 0$. The asymptotic loss can be computed analytically, and after tedious calculation, which can be found in the appendix, we find the following proposition.

Δ	θ	a	b_1
1	$\frac{3}{2}$	0.67 θ	0.005 Δ
$\frac{3}{2}$	1	0.765 θ	0.725 Δ
$\frac{3}{2}$	$\frac{3}{2}$	0.7 θ	0.595 Δ
$\frac{3}{2}$	2	0.67 θ	0.005 Δ
1	0	0	0.98 Δ
$\frac{3}{2}$	0	0	0.805 Δ

Table 5.2: Estimated values for a and b for SVM in the normal model, for several values of Δ and θ.

Proposition 5.7 *If $\theta = 0$, the asymptotic loss for support vector machines satisfies*

$$A\text{-Loss}_L(H_m) = \frac{p\Phi(T)}{4A_0^2 C_L(H_m)^2\Delta}\phi\Big(-\frac{\Delta}{2}\Big) \tag{5.13}$$

at the canonical model, where $\phi(\cdot)$ and $\Phi(\cdot)$ are the density and cumulative density function of the standard normal distribution, and where

$$T = \frac{1 - C_L(H_m)\Delta^2/2}{C_L(H_m)\Delta} \text{ and } A_0 = \frac{\phi(T)}{C_L(H_m)\Delta}.$$

5.4.4 Least squares

As mentioned above, we treat the least squares loss function differently because its loss function $L(u) = (u-1)^2$ is not monotone decreasing. The risk function to minimise is

$$R_{H_m}(a, b) = E_{H_m}[(Y(a + b^t X) - 1)^2].$$

The minimum is found where the first derivative is zero, thus

$$\begin{aligned}\frac{\partial R_{H_m}(a,b)}{\partial a} &= 2E_{H_m}[Y(Y(a+b^tX)-1)] \\ &= 2\pi_+ E_{H_+}[a+b^tX-1] - 2\pi_- E_{H_-}[-a-b^tX-1] \\ &= 2a + (\pi_+ - \pi_-)\Delta b^t e_1 - 2(\pi_+ - \pi_-) \\ &= 2a + (\pi_+ - \pi_-)\Delta b_1 - 2(\pi_+ - \pi_-) = 0,\end{aligned}$$

leads to $a = \frac{1}{2}(\pi_+ - \pi_-)(2 - \Delta b_1)$. Differentiating the expected risk with respect to the slope parameters b gives

$$\begin{aligned}&\frac{\partial R_{H_m}(a,b)}{\partial b} \\ &= 2E_{H_m}[Y(Y(a+b^tX)-1)X] \\ &= 2\pi_+ E_{H_+}[X(a+X^tb-1)] - 2\pi_- E_{H_-}[X(-a-X^tb-1)] \\ &= (\pi_+ - \pi_-)a\Delta e_1 - \Delta e_1 + 2\Big(I_p + \frac{\Delta^2}{4}\Big)b \\ &= \frac{1}{2}(\pi_+ - \pi_-)(\pi_+ - \pi_-)(2 - \Delta b^t e_1)\Delta e_1 - \Delta e_1 + 2\Big(I_p + \frac{\Delta^2}{4}e_1e_1^t\Big)b \\ &= \big((\pi_+ - \pi_-)^2 - 1\big)\Delta e_1 + 2\Big(I_p + \frac{\Delta^2}{4}\big(1 - (\pi_+ - \pi_-)^2\big)e_1e_1^t\Big)b = 0,\end{aligned}$$

from which follows that

$$b = \frac{1}{2}\Big(I_p + \frac{\Delta^2}{4}\big(1 - (\pi_+ - \pi_-)^2\big)e_1e_1^t\Big)^{-1}\big((\pi_+ - \pi_-)^2 - 1\big)\Delta e_1 \neq C_L(H_m)\Delta e_1.$$

For Fisher consistency to hold, the matrix $I_p + \frac{\Delta^2}{4}\big(1 - (\pi_+ - \pi_-)^2\big)e_1e_1^t$ must be the identity matrix, up to a constant factor. However, this is only true in the degenerate case when either $\pi_+ = 1$ or $\pi_- = 1$. Hence, convex risk minimisation using the least squares loss function is *not* Fisher-consistent.

5.5 Numerical results

In this section we visualise the asymptotic relative classification efficiencies of the various convex risk minimisation methods.

Figure 5.1 contains graphs of the asymptotic loss of the various convex risk minimisation methods we analysed, and of the Fisher rule as benchmark, with

different settings for every plot. In each plot, the solid line corresponds to Fisher's rule, the dashed line to AdaBoost, the dotted line to logistic regression, and the dash-dotted line (only in plot (a)) corresponds to the support vector machine. For (a), we set the log-odds ratio $\theta = 0$, the number of variables $p = 2$, and we varied Δ between 0 and 8 on the horizontal axis. By looking at the asymptotic losses, we observe that logistic regression is the best CRM method amongst those we studied, and this holds for all Δ. When the two groups are well-mixed, $\Delta \leq 2$, SVM is the least efficient, whereas for larger Δ, the exponential loss function of AdaBoost significantly deteriorates its efficiency compared to the other techniques. If we set $\theta = 1$, which corresponds to plot (b), we again observe that logistic regression is more efficient than AdaBoost, this for all values of Δ. Here we do not plot SVM, because it is not Fisher consistent when θ differs from zero. For the bottom plot we kept $\Delta = 2$ fixed, and varied θ between 0 and 4. Once again, we observe that logistic regression is more efficient than AdaBoost. We believe logistic regression performs better than the other CRM methods, in terms of efficiency, because logistic regression can be written as a conditional maximum likelihood, whereas this cannot be done for the other CRM techniques.

To investigate exactly how efficient the studied CRM techniques are, we constructed plots of the asymptotic relative classification efficiencies, see Figure 5.2. We have used the same setting as for the plots in Figure 5.1. For the case where $\theta = 0$ is fixed (a), we verify that logistic regression is the most efficient of the studied CRM methods, and we observe that indeed, AdaBoost is more efficient than SVM when $\Delta \leq 2$. Also, we find that all the studied CRM methods are highly efficient ($> 80\%$) when the two populations are well-mixed, that is when $\Delta < 2$. When we fix $\theta = 1$, plot (b), we again observe that logistic regression is more efficient than AdaBoost, and that the latter seems to have lost some of its efficiency (40% when $\Delta = 2$) compared to when $\theta = 0$. If we keep $\Delta = 2$ constant, and vary θ between 0 and 4, we observe that both logistic regression and AdaBoost have lost efficiency when the populations become more unbalanced. The drop in efficiency is larger for AdaBoost.

Next, we consider the asymptotic case where we increase the number of variables p towards infinity. As mentioned above, the asymptotic relative classifi-

cation efficiencies are in this case determined by the $ASV(B_2)$ of each decision rule. This is illustrated in Figure 5.3, which shows the ARCEs for the three cases above, but with the number of variables $p \to \infty$. As in the two previous figures, we varied Δ between 0 and 8 while keeping $\theta = 0$ in (a), fixed $\theta = 1$ for (b), and we varied θ between 0 and 4 while holding $\Delta = 2$ in (c). We arrive at similar conclusions as in the case where $p = 2$.

5.6 Conclusions

Convex risk optimisation methods are a class of broadly adaptable techniques for binary classification. We compared the efficiency of these techniques in the specific setting of two normally distributed populations with the same variance. We calculated the asymptotic loss of these techniques using influence functions, and made comparisons with the Fisher linear discriminant rule, which is optimal in this setting. We found that for two badly-separated groups, where the Mahalanobis distance between the group means is smaller than 2, the studied convex risk minimisation techniques have a reasonably good efficiency if the two classes are reasonably balanced (θ close to zero).

One interesting topic for further research is to extend this theory to more general settings, such as two normally distributed populations with unequal variances or cases with non-normal distributions. However, for these more general settings, the computations are too complex to perform analytically, and numerical approximations or Monte Carlo simulations are needed.

Another interesting topic for future research is to investigate whether the theory developed in this paper can be extended to compute classification efficiencies with the regularisation parameter in the convex risk minimisation kept constant at the population level.

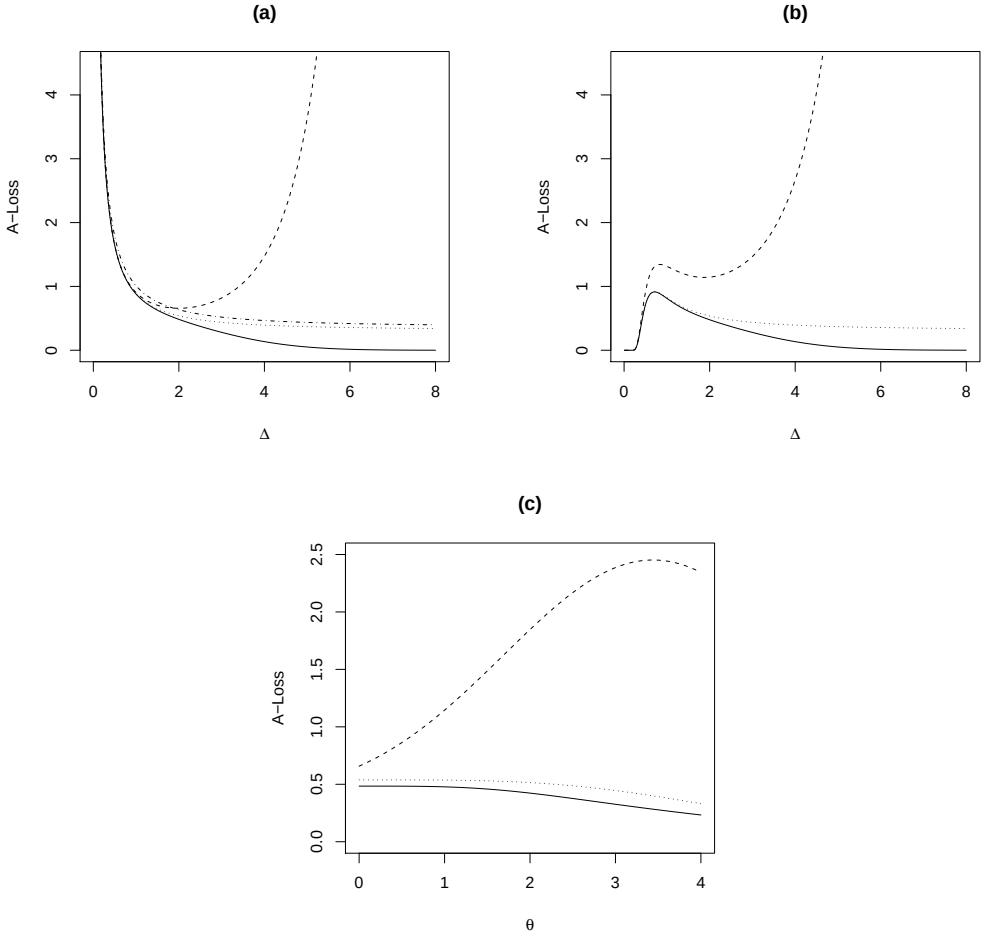

Figure 5.1: Asymptotic loss for Fisher's linear discriminant rule (solid), AdaBoost (dashed), logistic regression (dotted) and support vector machines (dash-dotted) with $p = 2$. (a) $\theta = 0$; (b) $\theta = 1$; (c) $\Delta = 1$.

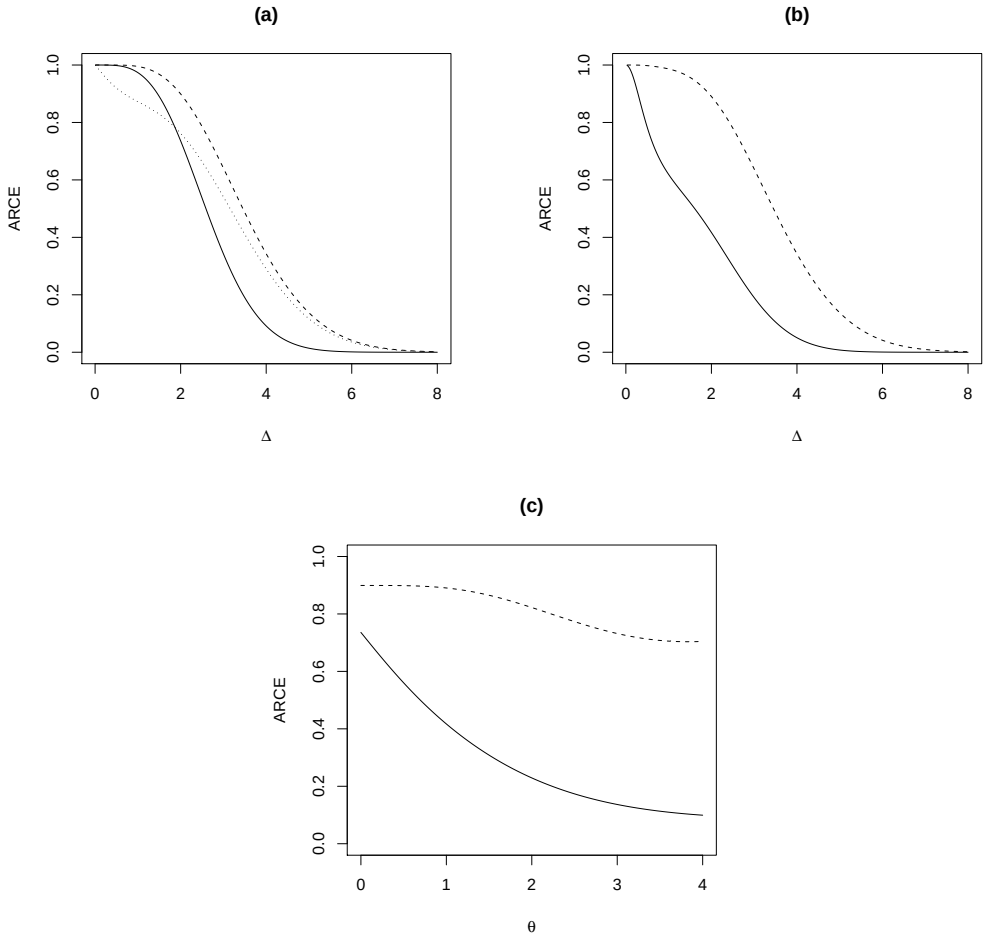

Figure 5.2: Asymptotic relative classification efficiencies for AdaBoost (solid), logistic regression (dashed), and SVM (dotted) for $p = 2$. (a) $\theta = 0$; (b) $\theta = 1$; (c) $\Delta = 1$.

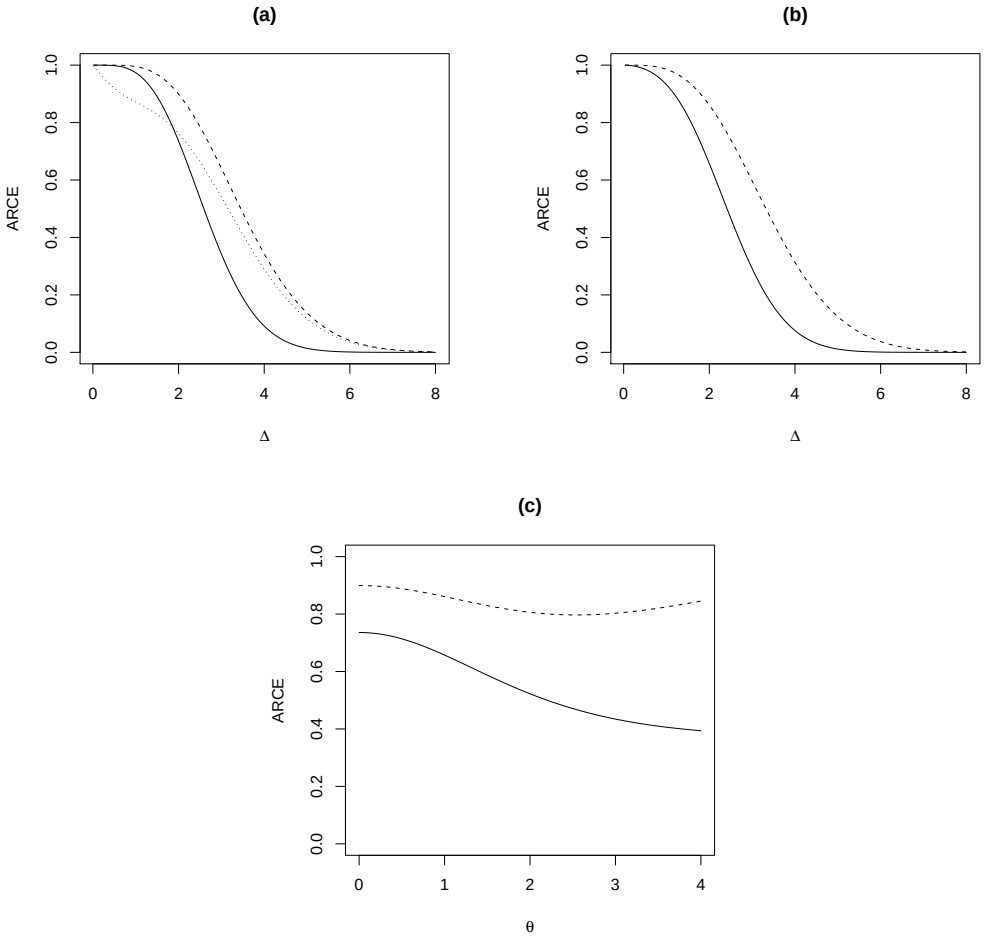

Figure 5.3: ARCEs for: AdaBoost (solid), logistic regression (dashed), and in (a), SVM (dotted), for $p = \infty$. (a) $\theta = 0$; (b) $\theta = 1$; (c) $\Delta = 1$.

Chapter 6

Discussion

In this thesis, we have addressed a few important issues in predictive modelling, ranging from focussed variable selection in the logistic regression model and the autoregressive time series model, where we provided several extensions to the existing focussed selection techniques, over variable selection in the support vector machine setting, to determining the classification efficiencies of various convex risk minimisation rules in the normal model. Most certainly, this thesis answers several questions, but it also raises several others.

The first question arises when we look at the boxplots of the error rates in Figure 2.2. Comparing the top plot to the bottom plot in that figure, we see that the median error rates decreases as the sample size n increases from 50 to 200 as expected. However, contrary to intuition, the highest observed generalisation error rates do not decrease with increasing sample size. Indeed, Efron (1975) demonstrated that, for a fixed model, the observed generalisation error of a linear decision rule for normally distributed populations with equal variance follows a χ^2 distribution, which explains the shape of the distribution of the observed generalisation error. For the error rates observed in our simulation experiment the situation is different, because the chosen model is not necessarily the same between each simulation run. Hence, we expect that the resulting distribution of estimated generalisation errors is a mixture of χ^2 distribution, arising from the

various different models, both good and bad ones. One topic for future research here is to verify the expectation, and to gain more insight into the exact nature of this mixture distribution.

We illustrated in Chapter 2 that using the focussed information criterion for variable selection results in a lower generalisation error for the selected models, due to the property that FIC can select different models for different new observations. However, this major strength of FIC can also be criticised, just because the user has to repeat the variable selection step for every new observation. Especially when a large number of predictions are needed, using this criterion becomes very time-consuming. Recently, Claeskens and Hjort (2008) developed a weighted version of the basic FIC based on mean squared error, which allows the user to select a model for an entire region of the space of observations. In the WESDR study addressed in Section 2.5, this would for example mean that it allows to select one model for all males living in an urban county, where the other variables are allowed to range freely over the entire observation space. A possible direction for future study is investigating the possibility of constructing weighted versions of the L_p-based FIC, and the FIC based on error rate, which would allow these criteria to be used for selecting a model for prediction across an entire region of the space of observations.

In Chapter 4, we have defined two new information criteria for support vector machines. We have chosen to let the penalty term depend linearly on the number of variables included in the model. This is not the only possible choice. An other alternative would be to use the (generalised) dimension of the feature (sub-)space as implied by the used kernel. The generalised dimension of the feature space would be akin to the degrees of freedom that a point has in that space. For the linear kernel, it will make no difference whether the number of variables or the dimension of the feature space is used. The advantage of using a penalty based on generalised dimension becomes apparent when using other kernels. With this penalty you acknowledge that more degrees of freedom are lost when a variable is eliminated while it can still be applied with the implied feature space of the Gaussian kernel, which has an infinite basis. The difference between both approaches, the number of variables and the generalised dimension of the input space, would

be an interesting topic to research.

A very important remark is that the criteria developed in Chapter 4 are meant to be used for selecting variables, and we did not concern ourselves with tuning the regularisation parameter C of the support vector machine (and by extension the kernel parameters). Nevertheless, it is well known that these parameters must be optimised as well to ensure that the support vector machine performs well as a classification tool. To mitigate this issue in some way, the user can optimise the hyperparameters of the support vector machine in the largest model under consideration, and use these during the variable selection step. Another good topic for future research is to examine whether these criteria can be used for simultaneous tuning of the hyperparameters and variable selection. Alternatively, it would be interesting to examine how these criteria perform if they are used for selection between models where the hyperparameters have been optimised for each model under consideration.

As mentioned in the conclusions to Chapter 4, the newly developed information criteria can also be used for variable selection in similar problems such as multicategory support vector machines or support vector regression. Especially in the latter case, where variable selection is very important, an examination of the performance of our proposed criteria is warranted.

We mentioned in Chapter 5 that we used the normal model because we knew the efficient rule in that setting, and also because it allowed us to find analytic expressions for the asymptotic losses of the studied decision rules. If this ideal situation does not hold, these two advantages disappear, and we are forced to obtain the asymptotic loss by means of simulation. Nonetheless, even though an efficient estimation method is not known in general, this technique would still allow us to compare the efficiency of two classification rules in a more general setting by comparing the losses with each other, without resorting to a benchmark efficient classification rule.

One important model which we haven't studied in Chapter 5 is the decision tree, which is also a very popular tool for classification, especially since it automatically performs variable selection. Our study did not include this technique because of the fact that decision trees cannot be interpreted as a convex risk

minimisation problem. Nevertheless, a comparison between decision trees and the convex risk minimisation methods could be interesting to explore in depth.

Appendix A

Proofs and computations

In this appendix we provide the proofs of the propositions, lemma, etc. We also give the analytical derivations for several results obtained in the main chapters. Appendix A.1 provides detailed results for the FIC based on L_p-norm from Chapter 2, Appendix A.2 gives the proofs for the results obtained in Chapter 3, and Appendix A.3 provides the proofs and analytical derivations for the results from Chapter 5.

A.1 FIC in logistic regression

Computation of the L_p-norm related risk $r_p(S)$, for p integer.

For $\Lambda_S \sim \mathcal{N}(\lambda, \sigma^2)$, we write $E[|\Lambda_S|^p] = E[|\sigma Z + \lambda|^p]$ where Z has a standard normal distribution. From this it follows that:

$$\begin{aligned} E[|\Lambda_S|^p] &= \frac{1}{\sqrt{2\pi}} \int_{-\frac{\lambda}{\sigma}}^{+\infty} (\sigma z + \lambda)^p \mathrm{e}^{-\frac{z^2}{2}} \,\mathrm{d}\, z + (-1)^p \int_{-\infty}^{-\frac{\lambda}{\sigma}} (\sigma z + \lambda)^p \mathrm{e}^{-\frac{z^2}{2}} \,\mathrm{d}\, z \\ &= \frac{1}{\sqrt{2\pi}} \sum_{j=0}^{p} \binom{p}{j} \sigma^j \lambda^{p-j} \left\{ \int_{-\frac{\lambda}{\sigma}}^{+\infty} z^j \mathrm{e}^{-\frac{z^2}{2}} \,\mathrm{d}\, z + (-1)^p \int_{-\infty}^{-\frac{\lambda}{\sigma}} z^j \mathrm{e}^{-\frac{z^2}{2}} \,\mathrm{d}\, z \right\}. \end{aligned}$$

From this expression, we can derive the following two formulae:

$$E[|\Lambda_S|^p] = \frac{1}{\sqrt{\pi}} \sum_{j'=0}^{p/2} \binom{p}{2j'} 2^{j'} \sigma^{2j'} \lambda^{p-2j'} \Gamma\left(j' + \frac{1}{2}\right),$$

for p even, and

$$\begin{aligned} E[|\Lambda_S|^p] &= \frac{1}{\sqrt{\pi}} \sum_{j'=0}^{(p-1)/2} \binom{p}{2j'} \sigma^{2j'} |\lambda|^{p-2j'} 2^{j'} \Gamma\left(j' + \frac{1}{2}\right) \\ &\quad + \frac{1}{\sqrt{\pi}} \sum_{j=0}^{p} \binom{p}{j} \sigma^j (-|\lambda|)^{p-j} 2^{j/2} \Gamma\left(\frac{j+1}{2}, \frac{\lambda^2}{2\sigma^2}\right), \end{aligned}$$

for p odd. Here, we denoted $\Gamma(\cdot)$ for the gamma function, and $\Gamma(a,x) = \int_x^{+\infty} t^{a-1} \mathrm{e}^{-t} \mathrm{d}\, t$ (for $a > 0$) for the incomplete gamma function.

For p even, say $p = 2r$, the expression can be simplified as follows.

$$\begin{aligned} E[|\Lambda_S|^{2r}] &= \frac{1}{\sqrt{2\pi}} \sum_{j=0}^{2r} \binom{2r}{j} \sigma^j \lambda^{2r-j} \int_{-\infty}^{+\infty} z^j \mathrm{e}^{-\frac{z^2}{2}} \,\mathrm{d}\, z \\ &= \sqrt{\frac{2}{\pi}} \sum_{j'=0}^{r} \binom{2r}{2j'} \sigma^{j'} \lambda^{2r-2j'} \int_0^{+\infty} z^{2j'} \mathrm{e}^{-\frac{z^2}{2}} \,\mathrm{d}\, z \\ &\overset{u=z^2/2}{=} \frac{1}{\sqrt{\pi}} \sum_{j'=0}^{r} \binom{2r}{2j'} 2^{j'} \sigma^{2j'} \lambda^{2r-2j'} \int_0^{+\infty} u^{j'-1/2} \mathrm{e}^{-u} \,\mathrm{d}\, u \\ &= \frac{1}{\sqrt{\pi}} \sum_{j'=0}^{r} \binom{2r}{2j'} 2^{j'} \sigma^{2j'} \lambda^{2r-2j'} \Gamma\left(j' + \frac{1}{2}\right). \end{aligned}$$

For p odd, say $p = 2r + 1$, this leads to

$$
\begin{aligned}
&E[|\Lambda_S|^p] \\
&= \frac{1}{\sqrt{2\pi}} \sum_{j=0}^{p} \binom{p}{j} \sigma^j \lambda^{p-j} \left\{ \int_{-\frac{\lambda}{\sigma}}^{+\infty} z^j \mathrm{e}^{-\frac{z^2}{2}} \,\mathrm{d}\, z - (-1)^j \int_{\frac{\lambda}{\sigma}}^{+\infty} z^j \mathrm{e}^{-\frac{z^2}{2}} \,\mathrm{d}\, z \right\} \\
&= \frac{1}{\sqrt{2\pi}} \sum_{j'=0}^{r} \left\{ \begin{array}{l} \binom{2r+1}{2j'} \sigma^{2j'} \lambda^{2r+1-2j'} \left\{ \int_{-\frac{\lambda}{\sigma}}^{+\infty} z^{2j'} \mathrm{e}^{-\frac{z^2}{2}} \,\mathrm{d}\, z - \int_{\frac{\lambda}{\sigma}}^{+\infty} z^{2j'} \mathrm{e}^{-\frac{z^2}{2}} \,\mathrm{d}\, z \right\} \\ + \binom{2r+1}{2j'+1} \sigma^{2j'+1} \lambda^{2r-2j'} \left\{ \int_{-\frac{\lambda}{\sigma}}^{+\infty} z^{2j'+1} \mathrm{e}^{-\frac{z^2}{2}} \,\mathrm{d}\, z + \int_{\frac{\lambda}{\sigma}}^{+\infty} z^{2j'+1} \mathrm{e}^{-\frac{z^2}{2}} \,\mathrm{d}\, z \right\} \end{array} \right\} \\
&= \sqrt{\frac{2}{\pi}} \sum_{j'=0}^{r} \left\{ \begin{array}{l} \binom{2r+1}{2j'} \sigma^{2j'} \lambda^{2r+1-2j'} \operatorname{sign}(\lambda) \int_0^{\frac{|\lambda|}{\sigma}} z^{2j'} \mathrm{e}^{-\frac{z^2}{2}} \,\mathrm{d}\, z \\ + \binom{2r+1}{2j'+1} \sigma^{2j'+1} \lambda^{2r-2j'} \int_{\frac{|\lambda|}{\sigma}}^{+\infty} z^{2j'+1} \mathrm{e}^{-\frac{z^2}{2}} \,\mathrm{d}\, z \end{array} \right\} \\
&\overset{u = z^2/2}{=} \frac{1}{\sqrt{\pi}} \sum_{j'=0}^{r} \left\{ \begin{array}{l} \binom{2r+1}{2j'} \sigma^{2j'} \lambda^{2r+1-2j'} \operatorname{sign}(\lambda) 2^{j'} \int_0^{\frac{\lambda^2}{2\sigma^2}} u^{j'-\frac{1}{2}} \mathrm{e}^{-u} \,\mathrm{d}\, u \\ + \binom{2r+1}{2j'+1} \sigma^{2j'+1} \lambda^{2r-2j'} 2^{j'+1/2} \int_{\frac{\lambda^2}{2\sigma^2}}^{+\infty} u^j \mathrm{e}^{-u} \,\mathrm{d}\, u \end{array} \right\} \\
&= \frac{1}{\sqrt{\pi}} \sum_{j'=0}^{r} \left\{ \begin{array}{l} \binom{2r+1}{2j'} \sigma^{2j'} \lambda^{2r+1-2j'} \operatorname{sign}(\lambda) 2^{j'} \{\Gamma(j'+\frac{1}{2}) - \Gamma(j'+\frac{1}{2}, \frac{\lambda^2}{2\sigma^2})\} \\ + \binom{2r+1}{2j'+1} \sigma^{2j'+1} \lambda^{2r-2j'} 2^{j'+1/2} \Gamma(j'+1, \frac{\lambda^2}{2\sigma^2}) \end{array} \right\} \\
&= \frac{1}{\sqrt{\pi}} \sum_{j'=0}^{r} \binom{2r+1}{2j'} \sigma^{2j'} |\lambda|^{2r+1-2j'} 2^{j'} \Gamma\left(j' + \frac{1}{2}\right) \\
&\qquad + \frac{1}{\sqrt{\pi}} \sum_{j=0}^{2r+1} \binom{2r+1}{j} \sigma^j (-|\lambda|)^{2r+1-j} 2^{j/2} \Gamma\left(\frac{j+1}{2}, \frac{\lambda^2}{2\sigma^2}\right).
\end{aligned}
$$

This ends the proof.

A.2 FIC for time series

Assumptions

We make the following assumptions on the series $\{x_t\}$ and $\{y_t\}$:

(A1) The maximum and minimum eigenvalues of $\boldsymbol{X}^t\boldsymbol{X}$ satisfy (for constants $B > 0$ and $b > 0$)

$$\lambda_{\max}(\boldsymbol{X}^t\boldsymbol{X}) \leq BT; \qquad \lambda_{\min}(\boldsymbol{X}^t\boldsymbol{X}) \geq bT,$$

where $\boldsymbol{X} = \big(\boldsymbol{x}_{p_T+h+1}(p_T, h), \ldots, \boldsymbol{x}_T(p_T, h)\big)^t$.

(A2) We define $\boldsymbol{\alpha}_t(p_T, h) = (\boldsymbol{X}^t\boldsymbol{X})^{-1/2}\boldsymbol{x}_t(p_T, h)$, for $t = p_T + h + 1, \ldots, T$. Then uniformly in t_1 and t_2,

$$\boldsymbol{\alpha}_{t_1}(p_T, h)^t\boldsymbol{\alpha}_{t_2}(p_T, h) = \mathcal{O}(p_T/T).$$

(A3) $\|\boldsymbol{y}(p_T)\| = \mathcal{O}(\sqrt{p_T})$, and $\max\{|\boldsymbol{x}_t(p_T, h)^t\boldsymbol{y}(p_T)| : t = p_T + h + 1, \ldots, T\} = \mathcal{O}(p_T\sqrt{\log T})$.

These assumptions on the time series $\{x_t\}$ have an intuitive explanation. Assumption (A1) amounts to having an empirical autocovariance matrix which is bounded for all lengths T, and for which the inverse exists and is bounded. (A2) states that there are no outlying observations of the time series, and (A3) limits the extent of the dependency between the series $\{x_t\}$ and $\{y_t\}$.

We first prove the following lemma, which is an adaptation of Theorem 3.2 in Portnoy (1985) for the setting in which we work.

Lemma A.1 *Under assumptions (A1), (A2), and (A3), and bounding condition* $p_T\sqrt{\log T}/T \to 0$ *for* $T \to \infty$*, the following result holds,*

$$\boldsymbol{y}(p_T)^t\big(\hat{\boldsymbol{\delta}}(p_T, h) - \boldsymbol{\delta}_{\text{true}}(p_T, h)\big)\left(\frac{1}{v\sigma}\right) \to_d \mathcal{N}(0, 1) \text{ for } T \to \infty$$

where $v^2 = \boldsymbol{y}(p_T)^t(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{y}(p_T)$ *and* σ^2 *is as in model* (3.2)*.*

The proof follows the same lines as the proof of Theorem 3.2 in Portnoy (1985).

Proof. Let $\boldsymbol{b}(p_T) = (\boldsymbol{X}^t\boldsymbol{X})^{-1/2}\boldsymbol{y}(p_T)$. Then we can write that

$$v^2 = \|\boldsymbol{b}(p_T)\|^2 \quad \text{and} \quad \boldsymbol{y}(p_T)^t\big(\hat{\boldsymbol{\delta}}(p_T, h) - \boldsymbol{\delta}_{\text{true}}(p_T, h)\big)\left(\frac{1}{v\sigma}\right) = \frac{\boldsymbol{b}(p_T)^t\hat{\boldsymbol{\theta}}}{\|\boldsymbol{b}(p_T)\|},$$

with $\hat{\boldsymbol{\theta}} = \frac{1}{\sigma}(\boldsymbol{X}^t\boldsymbol{X})^{1/2}\big(\hat{\boldsymbol{\delta}}(p_T, h) - \boldsymbol{\delta}_{\text{true}}(p_T, h)\big)$. It suffices to show that, for $\|\boldsymbol{b}(p_T)\| = 1$, $\boldsymbol{b}(p_T)^t\hat{\boldsymbol{\theta}} \to_d \mathcal{N}(0, 1)$. So assume that $\|\boldsymbol{b}(p_T)\| = 1$. For OLS estimation and normally distributed error terms, Lemma 3.4 of Portnoy (1985) is applicable, and gives

$$\boldsymbol{b}(p_T)^t\hat{\boldsymbol{\theta}} = \frac{1}{\sigma}\sum_{t=p_T+h}^{T}\boldsymbol{\alpha}_t(p_T, h)^t\boldsymbol{b}(p_T)\varepsilon_t(p_T, h).$$

Using the definition of $\boldsymbol{b}(p_T)$ and assumptions (A1) and (A2), we find $\boldsymbol{\alpha}_t(p_T,h)^t\boldsymbol{b}(p_T) = \boldsymbol{x}_t(p_T,h)^t(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{y}(p_T) = \frac{c}{T}\boldsymbol{x}_t(p_T,h)^t\boldsymbol{y}(p_T)$ for some constant c. Using assumption (A3) and the constraint on p_T, we then arrive at $\max_t|\boldsymbol{\alpha}_t(p_T,h)\boldsymbol{b}(p_T)| = \mathcal{O}(p_T\sqrt{T}/T) \to 0$ as $T\to\infty$. With

$$\sum_{t=p_T+h}^{T}\big(\boldsymbol{\alpha}_t(p_T,h)^t\boldsymbol{b}(p_T)\big)^2 = \|\boldsymbol{b}(p_T)\|^2 = 1,$$

the Central Limit Theorem implies that $\boldsymbol{b}(p_T)^t\hat{\boldsymbol{\theta}} \to_d \mathcal{N}(0,1)$ as $T\to\infty$, and the lemma holds. □

Proof of Proposition 3.1

Proof. For h-step ahead prediction,

$$\begin{aligned}
&\sqrt{T}\big(\hat{\mu}(p,h) - \mu_{\text{true}}(p_T,h)\big)\\
&\quad= \sqrt{T}\big(\hat{\mu}(p,h) - \hat{\mu}_{\text{true}}(p,h)\big) + \sqrt{T}\big(\hat{\mu}_{\text{true}}(p,h) - \hat{\mu}_{\text{true}}(p_T,h)\big)\\
&\quad= \sqrt{T}\big(\hat{\boldsymbol{\phi}}(p,h) - \boldsymbol{\phi}(p,h)\big)^t\boldsymbol{y}(p) + \sqrt{T}\big(\boldsymbol{\phi}(p,h)^t y(p) - \boldsymbol{\phi}(p_T,h)^t\boldsymbol{y}(p_T)\big).
\end{aligned}$$

The first term converges in distribution to a normal distribution. This follows by application of the limiting result in Hjort & Claeskens (2003, Lemma 3.3), where the maximal order is equal to p finite. The second term converges to a constant, since $\boldsymbol{\phi}(p_T,h)^t\boldsymbol{y}(p_T)$ is $\mathcal{O}_p(1/\sqrt{T})$. Hence, for each p fixed, the proposition holds.

However, the proposition must also hold for a growing number of time series components:

$$\begin{aligned}
\sqrt{T}\big(\mu(p_T,h) - \mu_{\text{true}}(p_T,h)\big) &= \sqrt{T}\big(\hat{\boldsymbol{\phi}}(p_T,h) - \boldsymbol{\phi}(p_T,h)\big)^t\boldsymbol{y}(p_T)\\
&= \big(\hat{\boldsymbol{\delta}}(p_T,h) - \boldsymbol{\delta}(p_T,h)\big)^t\boldsymbol{y}(p_T).
\end{aligned}$$

Lemma A.1 proves that this converges to a normal distribution as $T\to\infty$, and the proposition holds. □

Proof of Extension 3.6.1

Assume that h is the fixed prediction horizon and assume that p_T, the maximal AR-order of the considered models, satisfies the condition in Proposition 3.1. We

also assume that $h \leq p_T$. Then recursive substitution reveals that

$$\begin{aligned}\hat{\mu}(p_T,h) &= \hat{\phi}_1(p_T)\hat{\mu}(p_T,h-1)+\cdots+\hat{\phi}_{h-1}(p_T)\hat{\mu}(p_T,1)\\ &\quad+\hat{\phi}_h(p_T)y_T+\cdots+\hat{\phi}_{p_T}(p_T)y_{T+h-p_T}\\ &= \Big(\hat{\phi}_h(p_T)+\tilde{g}_1\big(\hat{\boldsymbol{\phi}}(p_T)\big)\Big)y_T+\cdots+\Big(\hat{\phi}_{p_T}(p_T)+\tilde{g}_{p_T-h}\big(\hat{\boldsymbol{\phi}}(p_T)\big)\Big)y_{T+h-p_T}\\ &\quad+\tilde{g}_{p_T-h+1}\big(\hat{\boldsymbol{\phi}}(p_T)\big)y_{T+h-p_T-1}+\cdots+\tilde{g}_{p_T}\big(\hat{\boldsymbol{\phi}}(p_T)\big)y_{T-p_T},\end{aligned}$$

where we used that $\hat{\mu}(p_T,-i)=y_{T-i}$ for $i\geq 0$. In this expression, $\tilde{g}_i\big(\hat{\boldsymbol{\phi}}(p_T)\big)$ for $1\leq i\leq p_T$ are polynomials of degree h in $\hat{\phi}_1(p_T),\ldots,\hat{\phi}_{p_T}(p_T)$ without a constant term or a first degree term. Since $\hat{\boldsymbol{\phi}}(p_T)=\hat{\boldsymbol{\delta}}(p_T)/\sqrt{T}$, it can be verified easily that $\tilde{g}_i\big(\hat{\boldsymbol{\phi}}(p_T)\big)=\mathcal{O}_p(1/T)$ for all $1\leq i\leq p_T$. We use this to rewrite the expression $\sqrt{T}(\hat{\mu}(p_T,h)-\mu(p_T,h))$ as

$$\sqrt{T}(\hat{\mu}(p_T,h)-\mu(p_T,h))=\sqrt{T}\Big(\sum_{i=0}^{p_T-h}\hat{\phi}_{h+i}(p_T)y_{T-i}\Big)+\sqrt{T}\Big(\sum_{i=0}^{p_T}\tilde{g}_i\big(\hat{\boldsymbol{\phi}}(p_T)\big)y_{T-i}\Big),$$

where $\mu(p_T,h)$ is the true value of the plug-in estimator. From the previous argument about the convergence rate of $\tilde{g}_i\big(\hat{\boldsymbol{\phi}}(p_T)\big)$, we see that the second term is $\mathcal{O}_p(1/\sqrt{T})$ and hence will have no contribution in the limit. We can then apply the same reasoning as in the proof of Proposition 3.1, but with $\tilde{\boldsymbol{y}}(p_T)=(0,\ldots,0,y_T,\ldots,y_{T+h-p_T})^t$ of length p_T, which proves the validity of Extension 3.6.1. □

A.3 CRM classification efficiencies

Proof of Proposition 5.2: Under the conditions listed in the proposition, and at the canonical model distribution, we have that

$$\begin{aligned}R_{L,H_m}(a,b) &= \frac{1}{2}E_{H_+}[L(a+b^tX)]+\frac{1}{2}E_{H_-}[L(-a-b^tX)]\\ &= \frac{1}{2}E\Big[L\Big(\frac{b_1\Delta}{2}+\sqrt{b^tb}Z+a\Big)\Big]+\frac{1}{2}E\Big[L\Big(\frac{b_1\Delta}{2}+\sqrt{b^tb}Z-a\Big)\Big],\end{aligned}$$

where Z follows a standard normal distribution and $b_1=b^te_1$. Denote $\sigma(b)=\sqrt{b^tb}$. We first keep $\sigma>0$ fixed and minimise $R_{L,H_m}(a,b)$ over b under the restriction that $\sigma(b)=\sigma$. Because $L(\cdot)$ is a decreasing function, it follows immediately

that $R_{L,H_m}(a,b)$ decreases as b_1 increases. The Cauchy-Schwarz inequality yields

$$b_1\Delta \leq \|b\|_2\Delta = \sigma\Delta,$$

and this inequality becomes an equality if $b = \sigma e_1$. Hence, the minimal value of the expected risk for a and σ fixed becomes

$$R_{L,H_m}(a,\sigma) = \frac{1}{2}E\Big[L\Big(\frac{\sigma\Delta}{2} + \sigma Z + a\Big)\Big] + \frac{1}{2}E\Big[L\Big(\frac{\sigma\Delta}{2} + \sigma Z - a\Big)\Big],$$

which has to be minimised with respect to σ and a.

Setting the first order derivative with respect to a to zero implies that the optimal σ and a have to satisfy

$$E\Big[L'\Big(\frac{\sigma\Delta}{2} + \sigma Z + a\Big)\Big] = E\Big[L'\Big(\frac{\sigma\Delta}{2} + \sigma Z - a\Big)\Big].$$

This can only be true if $a = 0$, from which it follows that $A_L(H_m) = 0$. Then, with $C_L(H_m)/\Delta$ the minimiser of $R_{L,H_m}(0,\sigma)$, one has

$$B_L(H_m) = C_L(H_m)\Delta e_1.$$

Since $\theta = 0$, we conclude that (5.7) holds, from which Fisher consistency follows. □

Proof of Proposition 5.3: Denote $\tilde{B}_L(H) = \big(A_L(H), B_L(H)^t\big)^t$, $\tilde{X} = (1, X^t)^t$, and $\tilde{x} = (1, x^t)^t$. We know that the first order derivatives of $R_{L,H}(a,b)$ with respect to a and b, evaluated in $\tilde{B}_L(H)$, are equal to zero. This holds in particular for $H_\varepsilon = (1-\varepsilon)H_m + \varepsilon\Delta_{\tilde{x}}$. With $\tilde{b}_\varepsilon = \tilde{B}_L(H_\varepsilon)$, we have that

$$\begin{aligned}
\psi\big(\varepsilon, \tilde{b}_\varepsilon\big) &\equiv \frac{\partial}{\partial\tilde{b}} R_{L,H_\varepsilon}(\tilde{b})\Big|_{\tilde{b}_\varepsilon} \\
&= (1-\varepsilon)\frac{\partial}{\partial\tilde{b}} E_{H_m}[L(Y\tilde{b}^t\tilde{X})]\Big|_{\tilde{b}_\varepsilon} + \varepsilon\frac{\partial}{\partial\tilde{b}} L(y\tilde{b}^t\tilde{x})\Big|_{\tilde{b}_\varepsilon} \\
&= (1-\varepsilon)E_{H_m}[YL'(Y\tilde{b}_\varepsilon^t\tilde{X})\tilde{X}] + \varepsilon yL'(y\tilde{b}_\varepsilon^t\tilde{x})\tilde{x} = 0
\end{aligned}$$

holds for all ε. From this it follows that

$$\frac{\mathrm{d}\,\psi(\varepsilon,\tilde{b}_\varepsilon)}{\mathrm{d}\,\varepsilon}\Big|_{\varepsilon=0} = \frac{\partial\psi(\varepsilon,\tilde{b}_0)}{\partial\varepsilon}\Big|_{\varepsilon=0} + \Big(\frac{\partial\psi\big(0,\tilde{b}_\varepsilon\big)}{\partial\tilde{b}}\Big|_{\tilde{b}=\tilde{b}_0}\Big)\Big(\frac{\partial\tilde{b}_\varepsilon}{\partial\varepsilon}\Big|_{\varepsilon=0}\Big) = 0,$$

or in other words, that

$$\begin{aligned}\frac{\partial\tilde{b}_\varepsilon}{\partial\varepsilon}\Big|_{\varepsilon=0} &= \mathrm{IF}\big((x,y);\tilde{B}_L,H_m\big)\\ &= -\Big(\frac{\partial\psi\big(0,\tilde{b}_\varepsilon\big)}{\partial\tilde{B}_L}\Big|_{\tilde{b}_\varepsilon=\tilde{B}_L(H_m)}\Big)^{-1}\Big(\frac{\partial\psi\big(\varepsilon,\tilde{B}_L(H_m)\big)}{\partial\varepsilon}\Big|_{\varepsilon=0}\Big). \quad \text{(A.1)}\end{aligned}$$

The second factor is easily seen to be

$$\begin{aligned}\frac{\partial\psi\big(\varepsilon,\tilde{B}_L(H_m)\big)}{\partial\varepsilon}\Big|_{\varepsilon=0} &= -E_{H_m}[YL'(Y\tilde{B}_L(H_m)^t\tilde{X})\tilde{X}] + yL'(y\tilde{B}_L(H_m)^t\tilde{x})\tilde{x}\\ &= yL'\big(yC_L(H_m)(\theta+\Delta x_1)\big)\tilde{x},\end{aligned}$$

because $\psi(0,\tilde{b}_0)=0$.
For the first factor, observe that

$$\begin{aligned}\frac{\partial\psi\big(0,\tilde{b}_\varepsilon\big)}{\partial\tilde{B}_L}\Big|_{\tilde{b}_\varepsilon=\tilde{B}_L(H_m)} &= E_{H_m}\big[L''\big(Y\tilde{B}_L(H_m)\tilde{X}\big)\tilde{X}\tilde{X}^t\big]\\ &= E_{H_m}\big[L''\big(YC_L(H_m)(\theta+\Delta X_1)\big)\tilde{X}\tilde{X}^t\big] = K.\end{aligned}$$

Because we work in the canonical model defined in Section 5.3, it immediately follows from symmetry arguments that

$$K = \left(\begin{array}{cc|c} A_0 & A_1 & 0\\ A_1 & A_2 & 0\\ \hline 0 & 0 & A_0 I_{p-1}\end{array}\right),$$

with A_0, A_1, and A_2 as in (5.8). From (A.1) we can easily find the expressions for the first-order influence functions as stated in proposition 5.3. □

Proof of Corollary 5.4: We have

$$\lim_{x_1\to-\infty}\mathrm{IF}\big((x,y);A_L,H_m\big) = -\lim_{x_1\to-\infty} yL'\big(yC_L(H_m)(\theta+\Delta x_1)\big)\frac{A_2-A_1x_1}{D}.$$

Because L is a decreasing, convex function, we find that L' is a negative, increasing function, where it is defined. Hence, it holds that

$$\lim_{x_1\to-\infty} L'\big(yC_L(H_m)(\theta+\Delta x_1)\big) < 0,$$

and

$$-\lim_{x_1\to-\infty} yL'\big(yC_L(H_m)(\theta+\Delta x_1)\big)\frac{A_2-A_1x_1}{D} = \text{sign}(y)\infty.$$

An analogous reasoning holds for the first order influence functions on the slope parameters. Thus, the first order influence functions on the parameters are unbounded, and by extension, the second order influence function on the error rate is unbounded as well. □

Proof of Proposition 5.5: To determine whether Fisher consistency holds for any θ, the values of c and d minimising

$$E_{H_m}\big[\exp\big(-Y(c\theta+d+c\Delta X_1)\big)\big]$$

must satisfy $d=0$, and $c=C_L(H_m)$.
We rewrite the expected risk as

$$\begin{aligned}
&E_{H_m}\big[\exp\big(-Y(c\theta+d+c\Delta X_1)\big)\big]\\
&= \pi_+E_{H_+}[\exp(-c\theta-d-c\Delta X_1)]+\pi_-E_{H_-}[\exp(c\theta+d+c\Delta X_1)]\\
&= \pi_+E\Big[\exp\Big(-c\theta-d-c\frac{\Delta^2}{2}-c\Delta Z\Big)\Big]+\pi_-E\Big[\exp\Big(c\theta+d-c\frac{\Delta^2}{2}+c\Delta Z\Big)\Big]\\
&= \pi_+\exp\Big(-c\theta-d-c\frac{\Delta^2}{2}+c^2\frac{\Delta^2}{2}\Big)+\pi_-\exp\Big(c\theta+d-c\frac{\Delta^2}{2}+c^2\frac{\Delta^2}{2}\Big).
\end{aligned}$$

The first order conditions with respect to c and d give

$$\begin{aligned}
0=\frac{\partial R_{L,H_m}(c,d)}{\partial d} &= -\pi_+\exp(-c\theta-d)\exp\Big(c^2\frac{\Delta^2}{2}-c\frac{\Delta^2}{2}\Big)\\
&\quad+\pi_-\exp(c\theta+d)\exp\Big(c^2\frac{\Delta^2}{2}-c\frac{\Delta^2}{2}\Big)
\end{aligned}$$

This is equivalent to $\pi_+\exp(-c\theta-d)=\pi_-\exp(c\theta+d)$, and thus $\pi_+/\pi_-=\exp(2c\theta+2d)$, and $\theta/2=c\theta+d$. Further,

$$\begin{aligned}
0 &= \frac{\partial R_{L,H_m}(c,d)}{\partial c}\\
&= \exp\Big(-c\frac{\Delta^2}{2}+c^2\frac{\Delta^2}{2}\Big)\Big(\pi_+\exp(-c\theta-d)\Big(-\theta-\frac{\Delta^2}{2}+c\Delta^2\Big)\\
&\qquad+\pi_-\exp(c\theta+d)\Big(\theta-\frac{\Delta^2}{2}+c\Delta^2\Big)\Big)
\end{aligned}$$

$$
\begin{aligned}
&= \pi_+ \exp\Big(-\frac{\theta}{2}\Big)\Big(-\theta-\frac{\Delta^2}{2}+c\Delta^2\Big)+\pi_- \exp\Big(\frac{\theta}{2}\Big)\Big(\theta-\frac{\Delta^2}{2}+c\Delta^2\Big)\\
&= \sqrt{\pi_+\pi_-}\Big(-\theta+\theta-\Delta^2+2c\Delta^2\Big).
\end{aligned}
$$

From this it follows that $c = 1/2$. Combined, this yields $d = 0$ and $c = C_L(H_m) = 1/2$. Hence, AdaBoost is Fisher consistent for all Δ and θ, which concludes the proof. □

Proof of Proposition 5.6: For computing the asymptotic loss, we first need to find an expression for the asymptotic variances of the parameters, see equation (5.9). The general form of these asymptotic parameters can be found in equation (5.10), and the unknown A_0, A_1, and A_2 are obtained using equation (5.8).

First, observe that we have $L''(u) = \exp(-u) = L(u)$. Using this result, we find that

$$
\begin{aligned}
A_n &= E_{H_m}\Big[\exp\Big(-\frac{Y}{2}(\theta+\Delta X_1)\Big)X_1^n\Big]\\
&= \pi_+ E_{H_+}\Big[\exp\Big(-\frac{Y}{2}(\theta+\Delta X_1)\Big)X_1^n\Big]+\pi_- E_{H_-}\Big[\exp\Big(\frac{Y}{2}(\theta+\Delta X_1)\Big)X_1^n\Big]\\
&= \Big(\pi_+\exp\Big(-\frac{\theta}{2}\Big)+(-1)^n\pi_-\exp\Big(\frac{\theta}{2}\Big)\Big)E_{H_+}\Big[\exp\Big(-\frac{\Delta}{2}X_1\Big)X_1^n\Big]\\
&= \begin{cases} 2\sqrt{\pi_+\pi_-}\exp\big(-\frac{\Delta^2}{4}\big)E\big[\exp\big(-\frac{\Delta}{2}Z\big)\big(\frac{\Delta}{2}+Z\big)^n\big] & \text{for } n \text{ even}\\ 0 & \text{for } n \text{ odd.}\end{cases}
\end{aligned}
$$

Using the above expression with $n = 1$, we obtain that $A_1 = 0$. In general, for $Z \sim \mathcal{N}(0,1)$ and any real c it holds that $E[\exp(cZ)] = \exp(c^2/2)$, $E[\exp(cZ)Z] = c\exp(c^2/2)$, and $E[\exp(cZ)Z^2] = (c^2+1)\exp(c^2/2)$. This is used to get that

$$
A_0 = 2\sqrt{\pi_+\pi_-}\exp\Big(-\frac{\Delta^2}{4}\Big)E\Big[\exp\Big(-\frac{\Delta}{2}Z\Big)\Big] = 2\sqrt{\pi_+\pi_-}\exp\Big(-\frac{\Delta^2}{8}\Big), \text{ and}
$$

$$
A_2 = 2\sqrt{\pi_+\pi_-}\exp\Big(-\frac{\Delta^2}{4}\Big)E\Big[\exp\Big(-\frac{\Delta}{2}Z\Big)\Big(\frac{\Delta}{2}+Z\Big)^2\Big] = 2\sqrt{\pi_+\pi_-}\exp\Big(-\frac{\Delta^2}{8}\Big).
$$

Having found A_0, A_1, and A_2, we compute the asymptotic variances in (5.10),

and we obtain

$$\begin{aligned}
\mathrm{ASV}(A) &= \frac{1}{4\pi_+\pi_-}\exp\Big(\frac{\Delta^2}{4}\Big)E_{H_m}\big[\exp\big(-Y(\theta+\Delta X_1)\big)\big] \\
&= \frac{1}{4\pi_+\pi_-}\exp\Big(\frac{\Delta^2}{4}\Big) \\
&\qquad \times\big(\pi_+E_{H_+}[\exp(-\theta-\Delta X_1)]+\pi_-E_{H_-}[\exp(\theta+\Delta X_1)]\big) \\
&= \frac{1}{4\pi_+\pi_-}\exp\Big(\frac{\Delta^2}{4}\Big) \\
&\qquad \times\big(\pi_+\exp(-\theta)+\pi_-\exp(\theta)\big)E\Big[\exp\Big(-\Delta\Big(\frac{\Delta}{2}+Z\Big)\Big)\Big] \\
&= \frac{1}{4\pi_+\pi_-}\exp\Big(\frac{\Delta^2}{4}\Big),
\end{aligned}$$

$$\begin{aligned}
\mathrm{ASV}(B_1) &= \frac{1}{4\pi_+\pi_-}\exp\Big(\frac{\Delta^2}{4}\Big)E_{H_m}\big[\exp\big(-Y(\theta+\Delta X_1)\big)X_1^2\big] \\
&= \frac{1}{4\pi_+\pi_-}\exp\Big(-\frac{\Delta^2}{4}\Big)E\Big[\exp(-\Delta Z)\Big(\frac{\Delta^2}{4}+\Delta Z+Z^2\Big)\Big] \\
&= \frac{1}{4\pi_+\pi_-}\exp\Big(\frac{\Delta^2}{4}\Big)\Big(\frac{\Delta^2}{4}+1\Big),
\end{aligned}$$

$$\begin{aligned}
\mathrm{ASV}(B_2) &= \frac{1}{4\pi_+\pi_-}\exp\Big(\frac{\Delta^2}{4}\Big)E_{H_m}\big[\exp\big(-Y(\theta+\Delta X_1)\big)X_2^2\big] \\
&= \frac{1}{4\pi_+\pi_-}\exp\Big(\frac{\Delta^2}{4}\Big)E_{H_m}\big[\exp\big(-Y(\theta+\Delta X_1)\big)\big]E_{H_m}[X_2^2] \\
&= \frac{1}{4\pi_+\pi_-}\exp\Big(\frac{\Delta^2}{4}\Big),
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{ASC}(A,B_1) &= \frac{1}{4\pi_+\pi_-}\exp\Big(\frac{\Delta^2}{4}\Big)E_{H_m}\big[\exp\big(-Y(\theta+\Delta X_1)\big)X_1\big] \\
&= \frac{1}{4\pi_+\pi_-}\exp\Big(-\frac{\Delta^2}{4}\Big)E\Big[\exp(-\Delta Z)\Big(\frac{\Delta}{2}+Z\Big)\Big] \\
&= -\frac{1}{8\pi_+\pi_-}\exp\Big(\frac{\Delta^2}{4}\Big).
\end{aligned}$$

Inserting the above quantities in (5.9), together with $C_L(H_m)=\frac{1}{2}$, yields the result in (5.12), proving the proposition. □

Proof of Proposition 5.7: Because we assume that $\theta = 0$, the expression in (5.9) simplifies to

$$\text{A-Loss}_L(H_m) = \frac{1}{4C_L(H_m)^2\Delta}\phi\Big(-\frac{\Delta}{2}\Big)\big(\text{ASV}(A) + (p-1)\text{ASV}(B_2)\big)$$

where $\text{ASV}(A)$ and $\text{ASV}(B_2)$ are defined as in (5.10), with $L'(u) = -I\{u \leq 1\}$ and $I\{\cdot\}$ the indicator function. For $Z \sim \mathcal{N}(0,1)$ and any real c, d it holds that $E[\delta(cZ-d)Z^j] = \big(\frac{d}{c}\big)^j\phi\big(\frac{d}{c}\big)/|c|$, for $j = 0, 1, 2$, where $\phi(\cdot)$ and $\Phi(\cdot)$ are respectively the density and the distribution function of the standard normal distribution, and $\delta(\cdot)$ is the Dirac-delta function. Using (5.8), and noting that $L''(u) = \delta(u-1)$, we evaluate the expressions for A_0, A_1, and A_2 as follows.

$$\begin{aligned}
A_0 &= E_{H_m}\big[\delta\big(YC_L(H_m)\Delta X_1 - 1\big)\big] \\
&= \frac{1}{2}E_{H_+}\big[\delta\big(C_L(H_m)\Delta X_1 - 1\big)\big] + \frac{1}{2}E_{H_-}\big[\delta\big(-C_L(H_m)\Delta X_1 - 1\big)\big] \\
&= \frac{1}{2}E\big[\delta\big(C_L(H_m)(\Delta^2/2 + \Delta Z) - 1\big)\big] \\
&\qquad + \frac{1}{2}E\big[\delta\big(-C_L(H_m)(-\Delta^2/2 + \Delta Z) - 1\big)\big] \\
&= \frac{1}{2C_L(H_m)\Delta}\phi\Big(\frac{1 - C_L(H_m)\Delta^2/2}{C_L(H_m)\Delta}\Big) + \frac{1}{2C_L(H_m)\Delta}\phi\Big(\frac{1 - C_L(H_m)\Delta^2/2}{C_L(H_m)\Delta}\Big) \\
&= \frac{1}{C_L(H_m)\Delta}\phi(T),
\end{aligned}$$

where we used the notation

$$T = \frac{1 - C_L(H_m)\Delta^2/2}{C_L(H_m)\Delta},$$

$$\begin{aligned}
A_1 &= E_{H_m}\big[\delta\big(YC_L(H_m)\Delta X_1 - 1\big)X_1\big] \\
&= \frac{1}{2}E_{H_+}\big[\delta\big(C_L(H_m)\Delta X_1 - 1\big)X_1\big] + \frac{1}{2}E_{H_-}\big[\delta\big(-C_L(H_m)\Delta X_1 - 1\big)X_1\big] \\
&= \frac{1}{2}E\big[\delta\big(C_L(H_m)(\Delta^2/2 + \Delta Z) - 1\big)(\Delta/2 + Z)\big] \\
&\qquad + \frac{1}{2}E\big[\delta\big(-C_L(H_m)(-\Delta^2/2 + \Delta Z) - 1\big)(-\Delta/2 + Z)\big] \\
&= \frac{\Delta\phi(T)}{4C_L(H_m)\Delta} + \frac{T\phi(T)}{2C_L(H_m)\Delta} - \frac{\Delta\phi(T)}{4C_L(H_m)\Delta} - \frac{T\phi(T)}{2C_L(H_m)\Delta} \\
&= 0,
\end{aligned}$$

and

$$\begin{aligned} A_2 &= E_{H_m}\big[\delta\big(YC_L(H_m)\Delta X_1 - 1\big)X_1^2\big] \\ &= \frac{1}{2}E\big[\delta\big(C_L(H_m)\Delta X_1 - 1\big)X_1^2\big] + \frac{1}{2}E\big[\delta\big(-C_L(H_m)\Delta X_1 - 1\big)X_1^2\big] \\ &= \frac{1}{2}E\big[\delta\big(C_L(H_m)(\Delta^2/2 + \Delta Z) - 1\big)(\Delta^2/4 + \Delta Z + Z^2)\big] \\ &\qquad + \frac{1}{2}E\big[\delta\big(-C_L(H_m)(-\Delta^2/2 + \Delta Z) - 1\big)(\Delta^2/4 - \Delta Z + Z^2)\big] \\ &= \frac{\phi(T)}{C_L(H_m)\Delta}\Big(\frac{\Delta^2}{4} + \Delta T + T^2\Big). \end{aligned}$$

For the evaluation of the asymptotic losses $\mathrm{ASV}(A)$ and $\mathrm{ASV}(B_2)$ we further use that for $Z \sim \mathcal{N}(0,1)$ and any real c, $E[I\{Z \le c\}] = \Phi(c)$, $E[I\{Z \le c\}Z] = -\phi(c)$, $E[I\{Z \le c\}Z^2] = \Phi(c) - c\phi(c)$, where $I\{\cdot\} = 1$ is the indicator function. This leads to

$$\begin{aligned} \mathrm{ASV}(A) &= \frac{1}{A_0}E_{H_m}[I\{YC_L(H_m)\Delta X_1 \le 1\}] \\ &= \frac{1}{2A_0^2}E_{+1}[I\{C_L(H_m)\Delta X_1 \le 1\}] + \frac{1}{2A_0^2}E_{-1}[I\{-C_L(H_m)\Delta X_1 \le 1\}] \\ &= \frac{1}{2A_0^2}E\Big[I\Big\{C_L(H_m)\Big(\frac{\Delta^2}{2} + \Delta Z\Big) \le 1\Big\}\Big] \\ &\qquad + \frac{1}{2A_0^2}E\Big[I\Big\{-C_L(H_m)\Big(-\frac{\Delta^2}{2} + \Delta Z\Big) \le 1\Big\}\Big] \\ &= \frac{\Phi(T)}{A_0^2} \end{aligned}$$

and $\mathrm{ASV}(B_2) = \mathrm{ASV}(A) = \Phi(T)/A_0^2$. Plugging these expressions into the asymptotic loss leads to

$$\text{A-Loss}_L(H_m) = \frac{1}{4C_L(H_m)^2\Delta}\phi\Big(-\frac{\Delta}{2}\Big)\Big(\frac{\Phi(T)}{A_0^2} + (p-1)\frac{\Phi(T)}{A_0^2}\Big),$$

which proves the proposition. □

List of Figures

List of Tables

Bibliography

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, in *Second International Symposium on Information Theory*, (eds. B. Petrov and F. Csáki), 267–281, Budapest: Akadémiai Kiadó.

Akaike, H. (1974). A new look at statistical model identification. *I.E.E.E. Transactions on Automatic Control*, **19**, 716–723.

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, **70**, 191–221.

Bhansali, R. J. (1996). Asymptotically efficient autoregressive model selection for multistep prediction. *Annals of the Institute of Statistical Mathematics*, **48**, 577–602.

Bi, J., Bennett, K. P., Embrechts, M., Breneman, C. M. and Song, M. (2003). Dimensionality reduction via sparse support vector machines. *Journal of Machine Learning Research*, **3**, 1229–1243.

Bollerslev, T. (1986). Generalised autoregressive conditional heteroskedasticity. *Journal of Econometrics*, **31**, 307–327.

Brockwell, P. J. and Davis, R. A. (1995). *Time series: Theory and methods*, 2nd edn. New York: Springer.

Burnham, K. P. and Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). New York: Springer.

Chen, S.-W., Li, Z.-R. and Li, X.-Y. (2005). Prediction of antifungal activity by support vector machine approach. *Journal of molecular structure: THEOCHEM*, **731**, 73–81.

Christmann, A., and Steinwart, I. (2004). On robustness properties of convex risk minimization methods for pattern recognition. *Journal of Machine Learning Research*, **5**, 1007–1034.

Claeskens, G., Croux, C. and Van Kerckhoven, J. (2006). Variable selection for logistic regression using a prediction focussed information criterion. *Biometrics*, **62**, 972–979.

Claeskens, G., Croux, C. and Van Kerckhoven, J. (2007). Prediction focussed model selection for autoregressive models. *Australian and New Zealand Journal of Statistics*, **49**, 359–379.

Claeskens, G. and Hjort, N. L. (2003). The focused information criterion [with discussion]. *Journal of the American Statistical Association*, **98**, 900–916.

Claeskens, G. and Hjort, N. L. (2008). *Model Selection and Model Averaging*. Cambridge: Cambridge University Press, in print.

Cristianini, N. and Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge: Cambridge University Press.

Croux, C., Filzmoser, P., and Joossens, K. (2008). Classification efficiencies for robust linear discriminant analysis. *Statistica Sinica*, to appear.

Croux, C., Haesbroeck, G., and Joossens, K. (2008). Logistic discrimination using robust estimators. *Canadian Journal of Statistics*, to appear.

Davis C. E., Hyde J. E., Bangdiwala S. I. and Nelson J. J. (1986). An example of dependencies among variables in a conditional logistic regression. *Modern Statistical Methods in Chronic Disease Epidemiology*, Eds. S.H. Moolgavkar and R.L. Prentice, New York: Wiley.

Diebold, F. X. (2001). *Elements of forecasting*, 2nd edn. Cincinatti (Ohio): South-Western Publ.

Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association*, **70**, 892–898.

Engle, R. E. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, **50**, 987–1007.

Fortuna, J. and Capson, D. (2004). Improved support vector classification using PCA and ICA feature space modification. *Pattern Recognition*, **37**, 1117–1129.

Freund, Y. and Schapire, R. (1996). Experiments with a new boosting algorithm. In *Machine Learning: Proceedings of the 13th International Conference*, pages 148–156. San Francisco: Morgan Kauffman Publishers.

Friedman, J., Hastie, T., and Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion). *Annals of Statistics*, **28**, 337–407.

George, E. I. (2000). The variable selection problem. *Journal of the American Statistical Association*, **95**, 1304–1308.

Golan, A., Judge, G. and Miller D. (1996). *Maximum entropy economics: robust estimation with limited data*. New York: John Wiley and Sons.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, **3**, 1157–1182.

Guyon, I., Gunn, S., Nikravesh, M., and Zadeh, L. (2006). *Feature Extraction: Foundations and Applications*. Berlin: Physica-Verlag, Springer.

Guyon, I., Li, J., Mader, T., Pletscher, P. A., Schneider G., and Uhr M. (2007). Competitive baseline methods set new standards for the NIPS 2003 feature selection benchmark. *Pattern Recognition Letters*, **28**, 1438–1444.

Guyon, I., Weston, J., Barnhill, S. and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, **46**, 389–422.

Hamilton, J. D. (1994). *Time series analysis*. Princeton: Princeton University Press.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust statistics: The approach based on influence functions*. New York: Wiley.

Hand, D. J. and Vinciotti, V. (2003). Local versus global models for classification problems: fitting models where it matters. *The American Statistician*, **57**, 124–131.

Hansen, B. E. (2005). Challenges for econometric model selection. *Econometric Theory*, **21**, 60–68.

Harvey, D. I., Leybourne, S. J. and Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, **13**, 281–291.

Harvey, D. I., Leybourne, S. J. and Newbold, P. (1998). Tests for forecast encompassing. *Journal of Business and Economic Statistics*, **16**, 254–259.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The elements of statistical learning: Data mining, inference and prediction*. New York: Springer.

Haughton, D. (1988). On the choice of a model to fit data from an exponential family, *The Annals of Statistics*, **16**, 342–355.

Haughton, D. (1989). Size of the error in the choice of a model to fit data from an exponential family, *Sankhyā, Series A*, **51**, 45–58.

Hjort, N. L. and Claeskens, G. (2003). Frequentist model average estimators [with discussion]. *Journal of the American Statistical Association*, **98**, 879–899.

Ing, C.-K. and Wei, C.-Z. (2005). Order selection for same-realization prediction in autoregressive processes. *Annals of Statistics*, **33**, 2423–2474.

Johnson, R. A. and Wichern, D. W. (1998). *Applied Multivariate Statistical Analysis*, 4th ed. New York: Prentic Hall.

Kearns, M., Mansour, Y., NG, A. Y. and Ron, D. (1997). An experimental and theoretical comparison of model selection methods. *Machine Learning*, **27**, 7–50.

Klein, R., Klein, B. E. K., Moss, S. E., Davis, M. D. and DeMets, D. L. (1984). The Wisconsin epidemiologic study of diabetic retinopathy: II. Prevalence and risk of diabetic retinopathy when age at diagnosis is less than 30 years, *Archives of Ophthalmology*, **102**, 520–526.

Kobayashi, K. and Komaki, F. (2006). Information criteria for support vector machines. *IEEE Transactions on Neural Networks*, **17**, 571–577.

Kuncheva, L. I. (2004). *Combining pattern classifiers: Methods and algorithms*. Hoboken (New Jersey): Wiley Interscience.

Le Cam, L. and Yang, G. L. (1990). *Asymptotics in statistics: Some basic concepts*. New York: Springer-Verlag.

Lee, S. and Karagrigoriou, A. (2001). An asymptotically optimal selection of the order of a linear process. *Sankhyā*, **63**, 93–106.

Lee, Y., Kim, Y., Lee, S., and Koo, J.-Y. (2006). Structured multicategory support vector machines with analysis of variance decomposition. *Biometrika*, **93**, 555–571.

Lin, Y. and Zhang, H. H. (2006). Component selection and smoothing in multivariate nonparametric regression. *Annals of Statistics*.

Ljung, G. M. and Box, G. E. P. (1979). The likelihood function of stationary autoregressive-moving average models. *Biometrika*, **66**, 265–270.

Neter, J., Kutner, M. H., Nachtsheim, C. J. and Wasserman, W. (1996). *Applied linear statistical models: Fourth edition*. Chicago (Illnois): Irwin.

Neumann, J., Schnörr, C. and Steidl, G. (2005). Combined SVM-based feature selection and classification. *Machine Learning*, **61**, 129–150.

Peng, H., Long, F. and Ding, C. (2005). Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE transactions on Pattern Analysis and Machine Intelligence*, **27**, 1226–1238.

Portnoy, S. (1985). Asymptotic behaviour of M estimators of p regression parameters when p^2/n is large; II. Normal approximation. *Annals of Statistics*, **13**, 1403–1417.

Rakotomamonjy, A. (2003). Variable selection using SVM-based criteria. *Journal of Machine Learning Research*, **3**, 1367–1370

Rätsch, G., Onoda, T. and Müller, K.-R. (2001). Soft margins for AdaBoost. *Machine Learning*, **42**, 287–320.

Rissanen, J. (1989). Stochastic complexity in statistical inquiry, *World Scientific Series in Computer Science*, volume 15. Singapore: World Scientific.

Schölkopf, B. and Smola, A. J. (2002). *Learning with kernels*. Cambridge (Massachusettes): MIT Press.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461–464.

Shi, P. and Tsai, C.-L. (2004). A joint regression variable and autoregressive order selection criterion. *Journal of Time Series Analysis*, **25**, 923–941.

Shibata, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Annals of Statistics*, **8**, 147–164.

Shih, F. Y. and Cheng, S. (2005). Improved feature reduction in input and feature spaces. *Pattern Recognition*, **38**, 651–659.

So, M. K. P., Chen, C. W. S. and Liu F.-C. (2006). Best subset selection of autoregressive models with exogenous variables and generalized autoregressive conditional heteroscedasticity errors. *Journal of the Royal Statistical Society C*, **55**, 201–224.

Sollich, P. (2002). Bayesian methods for support vector machines: Evidence and predictive class probabilities. *Machine Learning*, **46**, 21–52.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002). Bayesian measures of model complexity and fit [with discussion]. *Journal of the Royal Statistical Society B*, **64**, 583–639.

Van der Vaart (2000). *Asymptotic statistics*, Cambridge: Cambridge University Press.

Vapnik, V. N. (1982). *Estimation of dependences based on empirical data*. New York: Springer.

Vapnik, V. N. (1998). *Statistical learning theory*. New York: Wiley.

Wahba, G. (1999). Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV. In Schölkopf, B., Burges, C., Smola, A. J., editors, in *Advances in kernel methods – Support vector learning*, 69–88, Cambridge, (Massachusetts): MIT Press.

Woodroofe, M. (1982). On model selection and the arc sine laws. *The Annals of Statistics*, **10**, 1182–1194.

Zhang, H. H. (2006). Variable selection for SVM via smoothing spline ANOVA. *Statistica Sinica*, **16**, 659–674.

Zhang, T. (2004). Statistical behavior and consistency of classification methods based on convex risk minimisation. *The Annals of Statistics*, **32**, 56–85.

Zhang, X., Lu, X., Shi, Q., Xu, X.-Q., Leung, H.-C. E., Harris, L. N., Iglehart, J. D., Miron, A., Liu, J. S., and Wong, W. H. (2006). Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data. *BMC Bioinformatics*, published 10 April 2006.

Zhu, J., Rosset, S., Hastie, T. and Tibshirani, R. (2004). 1-norm support vector machines. *Neural Information Processing Systems*, **16**.

Doctoral Dissertations from the Faculty of Business and Economics

From August 1, 1971.

1. GEPTS Stefaan (1971)
   Stability and efficiency of resource allocation processes in discrete commodity spaces. Leuven, KUL, 1971. 86 pp.

2. PEETERS Theo (1971)
   Determinanten van de internationale handel in fabrikaten. Leuven, Acco, 1971. 290 pp.

3. VAN LOOY Wim (1971)
   Personeelsopleiding: een onderzoek naar investeringsaspekten van opleiding. Hasselt, Vereniging voor wetenschappelijk onderzoek in Limburg, 1971. VII, 238 pp.

4. THARAKAN Mathew (1972)
   Indian exports to the European community: problems and prospects. Leuven, Faculty of economics and applied economics, 1972. X,343 pp.

5. HERROELEN Willy (1972)
   Heuristische programmatie: methodologische benadering en praktische toepassing op complexe combinatorische problemen. Leuven, Aurelia scientifica, 1972. X, 367 pp.

6. VANDENBULCKE Jacques (1973)
   De studie en de evaluatie van data-organisatiemethodes en data-zoekmethodes. Leuven, s.n., 1973. 3 V.

7. PENNYCUICK Roy A. (1973)
   The economics of the ecological syndrome. Leuven, Acco, 1973. XII, 177 pp.

8. KAWATA T. Bualum (1973)
   Formation du capital d'origine belge, dette publique et stratégie du développement au Zaire. Leuven, KUL, 1973. V, 342 pp.

9. DONCKELS Rik (1974)
   Doelmatige oriëntering van de sectorale subsidiepolitiek in België: een theoretisch onderzoek met empirische toetsing. Leuven, K.U.Leuven, 1974. VII, 156 pp.

10. VERHELST Maurice (1974)
    Contribution to the analysis of organizational information systems and their financial benefits. Leuven, K.U.Leuven, 1974. 2 V.

11. CLEMEUR Hugo (1974)
    Enkele verzekeringstechnische vraagstukken in het licht van de nutstheorie. Leuven, Aurelia scientifica, 1974. 193 pp.

12. HEYVAERT Edward (1975)
    De ontwikkeling van de moderne bank- en krediettechniek tijdens de zestiende en zeventiende eeuw in Europa en te Amsterdam in het bijzonder. Leuven, K.U.Leuven, 1975. 186 pp.

13. VERTONGHEN Robert (1975)
    Investeringscriteria voor publieke investeringen: het uitwerken van een operationele theorie met een toepassing op de verkeersinfrastructuur. Leuven, Acco, 1975. 254 pp.

14. Niet toegekend.

15. VANOVERBEKE Lieven (1975)
    Microeconomisch onderzoek van de sectoriële arbeidsmobiliteit. Leuven, Acco, 1975. 205 pp.

16. DAEMS Herman (1975)
    The holding company: essays on financial intermediation, concentration and capital market imperfections in the Belgian economy. Leuven, K.U.Leuven, 1975. XII, 268 pp.

17. VAN ROMPUY Eric (1975)
    Groot-Brittannië en de Europese monetaire integratie: een onderzoek naar de gevolgen van de Britse toetreding op de geplande Europese monetaire unie. Leuven, Acco, 1975. XIII, 222 pp.

18. MOESEN Wim (1975)
Het beheer van de staatsschuld en de termijnstructuur van de intrestvoeten met een toepassing voor België. Leuven, Vander, 1975. XVI, 250 pp.

19. LAMBRECHT Marc (1976)
Capacity constrained multi-facility dynamic lot-size problem. Leuven, KUL, 1976. 165 pp.

20. RAYMAECKERS Erik (1976)
De mens in de onderneming en de theorie van het producenten-gedrag: een bijdrage tot transdisciplinaire analyse. Leuven, Acco, 1976. XIII, 538 pp.

21. TEJANO Albert (1976)
Econometric and input-output models in development planning: the case of the Philippines. Leuven, KUL, 1976. XX, 297 pp.

22. MARTENS Bernard (1977)
Prijsbeleid en inflatie met een toepassing op België. Leuven, KUL, 1977. IV, 253 pp.

23. VERHEIRSTRAETEN Albert (1977)
Geld, krediet en intrest in de Belgische financiële sector. Leuven, Acco, 1977. XXII, 377 pp.

24. GHEYSSENS Lieven (1977)
International diversification through the government bond market: a risk-return analysis. Leuven, s.n., 1977. 188 pp.

25. LEFEBVRE Chris (1977)
Boekhoudkundige verwerking en financiële verslaggeving van huurkooptransacties en verkopen op afbetaling bij ondernemingen die aan consumenten verkopen. Leuven, KUL, 1977. 228 pp.

26. KESENNE Stefan (1978)
Tijdsallocatie en vrijetijdsbesteding: een econometrisch onderzoek. Leuven, s.n., 1978. 163 pp.

27. VAN HERCK Gustaaf (1978)
Aspecten van optimaal bedrijfsbeleid volgens het marktwaardecriterium: een risico-rendements-analyse. Leuven, KUL, 1978. IV, 163 pp.

28. VAN POECK Andre (1979)
World price trends and price and wage development in Belgium: an investigation into the relevance of the Scandinavian model of inflation for Belgium. Leuven, s.n., 1979. XIV, 260 pp.

29. VOS Herman (1978)
De industriële technologieverwerving in Brazilië: een analyse. Leuven, s.n., 1978. onregelmatig gepagineerd.

30. DOMBRECHT Michel (1979)
Financial markets, employment and prices in open economies. Leuven, KUL, 1979. 182 pp.

31. DE PRIL Nelson (1979)
Bijdrage tot de actuariële studie van het bonus-malussysteem. Brussel, OAB, 1979. 112 pp.

32. CARRIN Guy (1979)
Economic aspects of social security: a public economics approach. Leuven, KUL, 1979. onregelmatig gepagineerd

33. REGIDOR Baldomero (1979)
An empirical investigation of the distribution of stock-market prices and weak-form efficiency of the Brussels stock exchange. Leuven, KUL, 1979. 214 pp.

34. DE GROOT Roger (1979)
Ongelijkheden voor stop loss premies gebaseerd op E.T. systemen in het kader van de veralgemeende convexe analyse. Leuven, KUL, 1979. 155 pp.

35. CEYSSENS Martin (1979)
On the peak load problem in the presence of rationizing by waiting. Leuven, KUL, 1979. IX, 217 pp.

36. ABDUL RAZK Abdul (1979)
Mixed enterprise in Malaysia: the case study of joint venture between Malysian public corporations and foreign enterprises. Leuven, KUL, 1979. 324 pp.

37. DE BRUYNE Guido (1980)
Coordination of economic policy: a game-theoretic approach. Leuven, KUL, 1980. 106 pp.

38. KELLES Gerard (1980)
Demand, supply, price change and trading volume on financial markets of the matching-order type. = Vraag, aanbod, koersontwikkeling en omzet op financiële markten van het Europese type. Leuven, KUL, 1980. 222 pp.

39. VAN EECKHOUDT Marc (1980)
De invloed van de looptijd, de coupon en de verwachte inflatie op het opbrengstverloop van vastrentende finaciële activa. Leuven, KUL, 1980. 294 pp.

40. SERCU Piet (1981)
Mean-variance asset pricing with deviations from purchasing power parity. Leuven, s.n., 1981. XIV, 273 pp.

41. DEQUAE Marie-Gemma (1981)
Inflatie, belastingsysteem en waarde van de onderneming. Leuven, KUL, 1981. 436 pp.

42. BRENNAN John (1982)
An empirical investigation of Belgian price regulation by prior notification: 1975 - 1979 - 1982. Leuven, KUL, 1982. XIII, 386 pp.

43. COLLA Annie (1982)
Een econometrische analyse van ziekenhuiszorgen. Leuven, KUL, 1982. 319 pp.

44. Niet toegekend.

45. SCHOKKAERT Eric (1982)
Modelling consumer preference formation. Leuven, KUL, 1982. VIII, 287 pp.

46. DEGADT Jan (1982)
Specificatie van een econometrisch model voor vervuilingsproblemen met proeven van toepassing op de waterverontreiniging in België. Leuven, s.n., 1982. 2 V.

47. LANJONG Mohammad Nasir (1983)
A study of market efficiency and risk-return relationships in the Malaysian capital market. s.l., s.n., 1983. XVI, 287 pp.

48. PROOST Stef (1983)
De allocatie van lokale publieke goederen in een economie met een centrale overheid en lokale overheden. Leuven, s.n., 1983. onregelmatig gepagineerd.

49. VAN HULLE Cynthia (1983)
Shareholders' unanimity and optimal corporate decision making in imperfect capital markets. s.l., s.n., 1983. 147 pp. + appendix.

50. VAN WOUWE Martine (2/12/83)
Ordening van risico's met toepassing op de berekening van ultieme ruïnekansen. Leuven, s.n., 1983. 109 pp.

51. D'ALCANTARA Gonzague (15/12/83)
SERENA: a macroeconomic sectoral regional and national account econometric model for the Belgian economy. Leuven, KUL, 1983. 595 pp.

52. D'HAVE Piet (24/02/84)
De vraag naar geld in België. Leuven, KUL, 1984. XI, 318 pp.

53. MAES Ivo (16/03/84)
The contribution of J.R. Hicks to macro-economic and monetary theory. Leuven, KUL, 1984. V, 224 pp.

54. SUBIANTO Bambang (13/09/84)
A study of the effects of specific taxes and subsidies on a firms' R&D investment plan. s.l., s.n., 1984. V, 284 pp.

55. SLEUWAEGEN Leo (26/10/84)
Location and investment decisions by multinational enterprises in Belgium and Europe. Leuven, KUL, 1984. XII, 247 pp.

56. GEYSKENS Erik (27/03/85)
Produktietheorie en dualiteit. Leuven, s.n., 1985. VII, 392 pp.

57. COLE Frank (26/06/85)
Some algorithms for geometric programming. Leuven, KUL, 1985. 166 pp.

58. STANDAERT Stan (26/09/86)
A study in the economics of repressed consumption. Leuven, KUL, 1986. X, 380 pp.

59. DELBEKE Jos (03/11/86)
Trendperioden in de geldhoeveelheid van België 1877-1983: een theoretische en empirische analyse van de "Banking school" hypothese. Leuven, KUL, 1986. XII, 430 pp.

60. VANTHIENEN Jan (08/12/86)
Automatiseringsaspecten van de specificatie, constructie en manipulatie van beslissingstabellen. Leuven, s.n., 1986. XIV, 378 pp.

61. LUYTEN Robert (30/04/87)
A systems-based approach for multi-echelon production/inventory systems. s.l., s.n., 1987. 3V.

62. MERCKEN Roger (27/04/87)
De invloed van de data base benadering op de interne controle. Leuven, s.n., 1987. XIII, 346 pp.

63. VAN CAYSEELE Patrick (20/05/87)
Regulation and international innovative activities in the pharmaceutical industry. s.l., s.n., 1987. XI, 169 pp.

64. FRANCOIS Pierre (21/09/87)
De empirische relevantie van de independence from irrelevant alternatives. Assumptie indiscrete keuzemodellen. Leuven, s.n., 1987. IX, 379 pp.

65. DECOSTER André (23/09/88)
Family size, welfare and public policy. Leuven, KUL. Faculteit Economische en Toegepaste Economische Wetenschappen, 1988. XIII, 444 pp.

66. HEIJNEN Bart (09/09/88)
Risicowijziging onder invloed van vrijstellingen en herverzekeringen: een theoretische analyse van optimaliteit en premiebepaling. Leuven, KUL. Faculteit Economische en Toegepaste Economische Wetenschappen, 1988. onregelmatig gepagineerd.

67. GEEROMS Hans (14/10/88)
Belastingvermijding. Theoretische analyse van de determinanten van de belastingontduiking en de belastingontwijking met empirische verificaties. Leuven, s.n., 1988. XIII, 409, 5 pp.

68. PUT Ferdi (19/12/88)
Introducing dynamic and temporal aspects in a conceptual (database) schema. Leuven, KUL. Faculteit Economische en Toegepaste Economische Wetenschappen, 1988. XVIII, 415 pp.

69. VAN ROMPUY Guido (13/01/89)
A supply-side approach to tax reform programs. Theory and empirical evidence for Belgium. Leuven, KUL. Faculteit Economische en Toegepaste Economische Wetenschappen, 1989. XVI, 189, 6 pp.

70. PEETERS Ludo (19/06/89)
Een ruimtelijk evenwichtsmodel van de graanmarkten in de E.G.: empirische specificatie en beleidstoepassingen. Leuven, K.U.Leuven. Faculteit Economische en Toegepaste Economische Wetenschappen, 1989. XVI, 412 pp.

71. PACOLET Jozef (10/11/89)
Marktstructuur en operationele efficiëntie in de Belgische financiële sector. Leuven, K.U.Leuven. Faculteit Economische en Toegepaste Economische Wetenschappen, 1989. XXII, 547 pp.

72. VANDEBROEK Martina (13/12/89)
Optimalisatie van verzekeringscontracten en premieberekeningsprincipes. Leuven, K.U.Leuven. Faculteit Economische en Toegepaste Economische Wetenschappen, 1989. 95 pp.

73. WILLEKENS Francois (1990)
Determinance of government growth in industrialized countries with applications to Belgium. Leuven, K.U.Leuven. Faculteit Economische en Toegepaste Economische Wetenschappen, 1990. VI, 332 pp.

74. VEUGELERS Reinhilde (02/04/90)
Scope decisions of multinational enterprises. Leuven, K.U.Leuven. Faculteit Economische en Toegepaste Economische Wetenschappen, 1990. V, 221 pp.

75. KESTELOOT Katrien (18/06/90)
Essays on performance diagnosis and tacit cooperation in international oligopolies. Leuven, K.U.Leuven. Faculteit Economische en Toegepaste Economische Wetenschappen, 1990. 227 pp.

76. WU Changqi (23/10/90) Strategic aspects of oligopolistic vertical integration. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1990. VIII, 222 pp.

77. ZHANG Zhaoyong (08/07/91)
A disequilibrium model of China's foreign trade behaviour. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1991. XII, 256 pp.

78. DHAENE Jan (25/11/91)
Verdelingsfuncties, benaderingen en foutengrenzen van stochastische grootheden geassocieerd aan verzekeringspolissen en -portefeuilles. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1991. 146 pp.

79. BAUWELINCKX Thierry (07/01/92)
Hierarchical credibility techniques. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1992. 130 pp.

80. DEMEULEMEESTER Erik (23/3/92)
Optimal algorithms for various classes of multiple resource-constrained project scheduling problems. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1992. 180 pp.

81. STEENACKERS Anna (1/10/92)
Risk analysis with the classical actuarial risk model: theoretical extensions and applications to Reinsurance. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1992. 139 pp.

82. COCKX Bart (24/09/92)
The minimum income guarantee. Some views from a dynamic perspective. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1992. XVII, 401 pp.

83. MEYERMANS Eric (06/11/92)
Econometric allocation systems for the foreign exchange market: Specification,

estimation and testing of transmission mechanisms under currency substitution. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1992. XVIII, 343 pp.

84. CHEN Guoqing (04/12/92)
Design of fuzzy relational databases based on fuzzy functional dependency. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1992. 176 pp.

85. CLAEYS Christel (18/02/93)
Vertical and horizontal category structures in consumer decision making: The nature of product hierarchies and the effect of brand typicality. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1993. 348 pp.

86. CHEN Shaoxiang (25/03/93)
The optimal monitoring policies for some stochastic and dynamic production processes. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1993. 170 pp.

87. OVERWEG Dirk (23/04/93)
Approximate parametric analysis and study of cost capacity management of computer configurations. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1993. 270 pp.

88. DEWACHTER Hans (22/06/93)
Nonlinearities in speculative prices: The existence and persistence of nonlinearity in foreign exchange rates. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1993. 151 pp.

89. LIN Liangqi (05/07/93)
Economic determinants of voluntary accounting choices for R & D expenditures in Belgium. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1993. 192 pp.

90. DHAENE Geert (09/07/93)
Encompassing: formulation, properties and testing. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1993. 117 pp.

91. LAGAE Wim (20/09/93)
Marktconforme verlichting van soevereine buitenlandse schuld door private crediteuren: een neo-institutionele analyse. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1993. 241 pp.

92. VAN DE GAER Dirk (27/09/93)
Equality of opportunity and investment in human capital. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1993. 172 pp.

93. SCHROYEN Alfred (28/02/94)
Essays on redistributive taxation when monitoring is costly. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1994. 203 pp. + V.

94. STEURS Geert (15/07/94)
Spillovers and cooperation in research and development. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1994. 266 pp.

95. BARAS Johan (15/09/94)
The small sample distribution of the Wald, Lagrange multiplier and likelihood ratio tests for homogeneity and symmetry in demand analysis: a Monte Carlo study. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1994. 169 pp.

96. GAEREMYNCK Ann (08/09/94)
The use of depreciation in accounting as a signalling device. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1994. 232 pp.

97. BETTENDORF Leon (22/09/94)
A dynamic applied general equilibrium model for a small open economy. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1994. 149 pp.

98. TEUNEN Marleen (10/11/94)
Evaluation of interest randomness in actuarial quantities. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1994. 214 pp.

99. VAN OOTEGEM Luc (17/01/95)
An economic theory of private donations. Leuven. K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1995. 236 pp.

100. DE SCHEPPER Ann (20/03/95)
Stochastic interest rates and the probabilistic behaviour of actuarial functions. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1995. 211 pp.

101. LAUWERS Luc (13/06/95)
Social choice with infinite populations. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1995. 79 pp.

102. WU Guang (27/06/95)
A systematic approach to object-oriented business modeling. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1995. 248 pp.

103. WU Xueping (21/08/95)
Term structures in the Belgian market: model estimation and pricing error analysis. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1995. 133 pp.

104. PEPERMANS Guido (30/08/95)
Four essays on retirement from the labor force. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1995. 128 pp.

105. ALGOED Koen (11/09/95)
Essays on insurance: a view from a dynamic perspective. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1995. 136 pp.

106. DEGRYSE Hans (10/10/95)
Essays on financial intermediation, product differentiation, and market structure. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1995. 218 pp.

107. MEIR Jos (05/12/95)
Het strategisch groepsconcept toegepast op de Belgische financiële sector. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1995. 257 pp.

108. WIJAYA Miryam Lilian (08/01/96)
Voluntary reciprocity as an informal social insurance mechanism: a game theoretic approach. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1996. 124 pp.

109. VANDAELE Nico (12/02/96)
The impact of lot sizing on queueing delays: multi product, multi machine models. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1996. 243 pp.

110. GIELENS Geert (27/02/96)
Some essays on discrete time target zones and their tails. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1996. 131 pp.

111. GUILLAUME Dominique (20/03/96)
Chaos, randomness and order in the foreign exchange markets. Essays on the modelling of the markets. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1996. 171 pp.

112. DEWIT Gerda (03/06/96)
Essays on export insurance subsidization. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1996. 186 pp.

113. VAN DEN ACKER Carine (08/07/96)
Belief-function theory and its application to the modeling of uncertainty in financial statement auditing. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1996. 147 pp.

114. IMAM Mahmood Osman (31/07/96)
Choice of IPO Flotation Methods in Belgium in an Asymmetric Information Framework and Pricing of IPO's in the Long-Run. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1996. 221 pp.

115. NICAISE Ides (06/09/96)
Poverty and Human Capital. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1996. 209 pp.

116. EYCKMANS Johan (18/09/97)
On the Incentives of Nations to Join International Environmental Agreements. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1997. XV + 348 pp.

117. CRISOLOGO-MENDOZA Lorelei (16/10/97)
Essays on Decision Making in Rural Households: a study of three villages in the Cordillera Region of the Philippines. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1997. 256 pp.

118. DE REYCK Bert (26/01/98)
Scheduling Projects with Generalized Precedence Relations: Exact and Heuristic Procedures. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1998. XXIV + 337 pp.

119. VANDEMAELE Sigrid (30/04/98)
Determinants of Issue Procedure Choice within the Context of the French IPO Market: Analysis within an Asymmetric Information Framework. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1998. 241 pp.

120. VERGAUWEN Filip (30/04/98)
Firm Efficiency and Compensation Schemes for the Management of Innovative Activities and Knowledge Transfers. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1998. VIII + 175 pp.

121. LEEMANS Herlinde (29/05/98)
The Two-Class Two-Server Queueing Model with Nonpreemptive Heterogeneous

Priority Structures. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1998. 211 pp.

122. GEYSKENS Inge (4/09/98)
Trust, Satisfaction, and Equity in Marketing Channel Relationships. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1998. 202 pp.

123. SWEENEY John (19/10/98)
Why Hold a Job ? The Labour Market Choice of the Low-Skilled. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1998. 278 pp.

124. GOEDHUYS Micheline (17/03/99)
Industrial Organisation in Developing Countries, Evidence from Côte d'Ivoire. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1999. 251 pp.

125. POELS Geert (16/04/99)
On the Formal Aspects of the Measurement of Object-Oriented Software Specifications. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1999. 507 pp.

126. MAYERES Inge (25/05/99)
The Control of Transport Externalities: A General Equilibrium Analysis. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1999. XIV + 294 pp.

127. LEMAHIEU Wilfried (5/07/99)
Improved Navigation and Maintenance through an Object-Oriented Approach to Hypermedia Modelling. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1999. 284 pp.

128. VAN PUYENBROECK Tom (8/07/99)
Informational Aspects of Fiscal Federalism. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1999. 192 pp.

129. VAN DEN POEL Dirk (5/08/99)
Response Modeling for Database Marketing Using Binary Classification. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1999. 342 pp.

130. GIELENS Katrijn (27/08/99)
International Entry Decisions in the Retailing Industry: Antecedents and Perfor-

mance Consequences. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1999. 336 pp.

131. PEETERS Anneleen (16/12/99)
Labour Turnover Costs, Employment and Temporary Work. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1999. 207 pp.

132. VANHOENACKER Jurgen (17/12/99)
Formalizing a Knowledge Management Architecture Meta-Model for Integrated Business Process Management. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 1999. 252 pp.

133. NUNES Paulo (20/03/2000)
Contingent Valuation of the Benefits of Natural Areas and its Warmglow Component. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2000. XXI + 282 pp.

134. VAN DEN CRUYCE Bart (7/04/2000)
Statistische discriminatie van allochtonen op jobmarkten met rigide lonen. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2000. XXIII + 441 pp.

135. REPKINE Alexandre (15/03/2000)
Industrial restructuring in countries of Central and Eastern Europe: Combining branch-, firm- and product-level data for a better understanding of Enterprises' behaviour during transition towards market economy. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2000. VI + 147 pp.

136. AKSOY, Yunus (21/06/2000)
Essays on international price rigidities and exchange rates. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2000. IX + 236 pp.

137. RIYANTO, Yohanes Eko (22/06/2000)
Essays on the internal and external delegation of authority in firms. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2000. VIII + 280 pp.

138. HUYGHEBAERT, Nancy (20/12/2000)
The Capital Structure of Business Start-ups. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2000. VIII + 332 pp.

139. FRANCKX Laurent (22/01/2001)
Ambient Inspections and Commitment in Environmental Enforcement. Leuven,

K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2001. VIII + 286 pp.

140. VANDILLE Guy (16/02/2001)
Essays on the Impact of Income Redistribution on Trade. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2001. VIII + 176 pp.

141. MARQUERING Wessel (27/04/2001)
Modeling and Forecasting Stock Market Returns and Volatility. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2001. V + 267 pp.

142. FAGGIO Giulia (07/05/2001)
Labor Market Adjustment and Enterprise Behavior in Transition. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2001. 150 pp.

143. GOOS Peter (30/05/2001)
The Optimal Design of Blocked and Split-plot experiments. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2001. X + 224 pp.

144. LABRO Eva (01/06/2001)
Total Cost of Ownership Supplier Selection based on Activity Based Costing and Mathematical Programming. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2001. 217 pp.

145. VANHOUCKE Mario (07/06/2001)
Exact Algorithms for various Types of Project Scheduling Problems. Nonregular Objectives and time/cost Trade-offs. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2001. 316 pp.

146. BILSEN Valentijn (28/08/2001)
Entrepreneurship and Private Sector Development in Central European Transition Countries. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2001. XVI + 188 pp.

147. NIJS Vincent (10/08/2001)
Essays on the dynamic Category-level Impact of Price promotions. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2001.

148. CHERCHYE Laurens (24/09/2001)
Topics in Non-parametric Production and Efficiency Analysis. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2001. VII + 169 pp.

149. VAN DENDER Kurt (15/10/2001)
Aspects of Congestion Pricing for Urban Transport. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2001. VII + 203 pp.

150. CAPEAU Bart (26/10/2001)
In defence of the excess demand approach to poor peasants' economic behaviour. Theory and Empirics of non-recursive agricultural household modelling. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2001. XIII + 286 pp.

151. CALTHROP Edward (09/11/2001)
Essays in urban transport economics. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2001.

152. VANDER BAUWHEDE Heidi (03/12/2001)
Earnings management in an Non-Anglo-Saxon environment. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2001. 408 pp.

153. DE BACKER Koenraad (22/01/2002)
Multinational firms and industry dynamics in host countries : the case of Belgium. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2002. VII + 165 pp.

154. BOUWEN Jan (08/02/2002)
Transactive memory in operational workgroups. Concept elaboration and case study. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2002. 319 pp. + appendix 102 pp.

155. VAN DEN BRANDE Inge (13/03/2002)
The psychological contract between employer and employee : a survey among Flemish employees. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2002. VIII + 470 pp.

156. VEESTRAETEN Dirk (19/04/2002)
Asset Price Dynamics under Announced Policy Switching. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2002. 176 pp.

157. PEETERS Marc (16/05/2002)
One Dimensional Cutting and Packing : New Problems and Algorithms. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2002. IX + 247 pp.

158. SKUDELNY Frauke (21/05/2002)
Essays on The Economic Consequences of the European Monetary Union. Leuven,

K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2002.

159. DE WEERDT Joachim (07/06/2002)
Social Networks, Transfers and Insurance in Developing countries. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2002. VI + 129 pp.

160. TACK Lieven (25/06/2002)
Optimal Run Orders in Design of Experiments. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2002. XXXI + 344 pp.

161. POELMANS Stephan (10/07/2002)
Making Workflow Systems work. An investigation into the Importance of Task-appropriation fit, End-user Support and other Technological Characteristics. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2002. 237 pp.

162. JANS Raf (26/09/2002)
Capacitated Lot Sizing Problems : New Applications, Formulations and Algorithms. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2002.

163. VIAENE Stijn (25/10/2002)
Learning to Detect Fraud from enriched Insurance Claims Data (Context, Theory and Applications). Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2002. 315 pp.

164. AYALEW Tekabe (08/11/2002)
Inequality and Capital Investment in a Subsistence Economy.Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2002. V + 148 pp.

165. MUES Christophe (12/11/2002)
On the Use of Decision Tables and Diagrams in Knowledge Modeling and Verification. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2002. 222 pp.

166. BROCK Ellen (13/03/2003)
The Impact of International Trade on European Labour Markets. K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2002.

167. VERMEULEN Frederic (29/11/2002)
Essays on the collective Approach to Household Labour Supply. Leuven, K.U.Leuven,

Faculteit Economische en Toegepaste Economische Wetenschappen, 2002. XIV + 203 pp.

168. CLUDTS Stephan (11/12/2002)
Combining participation in decision-making with financial participation : theoretical and empirical perspectives. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2002. XIV + 247 pp.

169. WARZYNSKI Frederic (09/01/2003)
The dynamic effect of competition on price cost margins and innovation. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2003.

170. VERWIMP Philip (14/01/2003)
Development and genocide in Rwanda ; a political economy analysis of peasants and power under the Habyarimana regime. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2003.

171. BIGANO Andrea (25/02/2003)
Environmental regulation of the electricity sector in a European Market Framework. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2003. XX + 310 pp.

172. MAES Konstantijn (24/03/2003)
Modeling the Term Structure of Interest Rates Across Countries. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2003. V+246 pp.

173. VINAIMONT Tom (26/02/2003)
The performance of One- versus Two-Factor Models of the Term Structure of Interest Rates. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen. 2003.

174. OOGHE Erwin (15/04/2003)
Essays in multi-dimensional social choice. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2003. VIII+108 pp.

175. FORRIER Anneleen (25/04/2003)
Temporary employment, employability and training. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2003.

176. CARDINAELS Eddy (28/04/2003)
The role of cost system accuracy in managerial decision making. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2003. 144 pp.

177. DE GOEIJ Peter (02/07/2003)
Modeling Time-Varying Volatility and Interest Rates. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2003. VII+225 pp.

178. LEUS Roel (19/09/2003)
The generation of stable project plans. Complexity and exact algorithms. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2003.

179. MARINHEIRO Carlos (23/09/2003)
EMU and fiscal stabilisation policy : the case of small countries. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2003.

180. BAESSENS Bart (24/09/2003)
Developing intelligent systems for credit scoring using machine learning techniques. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2003.

181. KOCZY Laszlo (18/09/2003)
Solution concepts and outsider behaviour in coalition formation games. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2003.

182. ALTOMONTE Carlo (25/09/2003)
Essays on Foreign Direct Investment in transition countries : learning from the evidence. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2003.

183. DRIES Liesbeth (10/11/2003)
Transition, Globalisation and Sectoral Restructuring: Theory and Evidence from the Polish Agri-Food Sector. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2003.

184. DEVOOGHT Kurt (18/11/2003)
Essays On Responsibility-Sensitive Egalitarianism and the Measurement of Income Inequality. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2003.

185. DELEERSNYDER Barbara (28/11/2003)
Marketing in Turbulent Times. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2003.

186. ALI Daniel (19/12/2003)
Essays on Household Consumption and Production Decisions under Uncertainty

in Rural Ethiopia. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2003.

187. WILLEMS Bert (14/01/2004)
Electricity networks and generation market power. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2004.

188. JANSSENS Gust (30/01/2004)
Advanced Modelling of Conditional Volatility and Correlation in Financial Markets. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2004.

189. THOEN Vincent (19/01/2004)
On the valuation and disclosure practices implemented by venture capital providers. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2004.

190. MARTENS Jurgen (16/02/2004)
A fuzzy set and stochastic system theoretic technique to validate simulation models. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2004.

191. ALTAVILLA Carlo (21/05/2004)
Monetary policy implementation and transmission mechanisms in the Euro area. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2004.

192. DE BRUYNE Karolien (07/06/2004)
Essays in the location of economic activity. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2004.

193. ADEM Jan (25/06/2004)
Mathematical programming approaches for the supervised classification problem. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2004.

194. LEROUGE Davy (08/07/2004)
Predicting Product Preferences : the effect of internal and external cues. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2004.

195. VANDENBROECK Katleen (16/07/2004)
Essays on output growth, social learning and land allocation in agriculture : microevidence from Ethiopia and Tanzania. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2004.

196. GRIMALDI Maria (03/09/2004)
The exchange rate, heterogeneity of agents and bounded rationality. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2004.

197. SMEDTS Kristien (26/10/2004)
Financial integration in EMU in the framework of the no-arbitrage theory. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2004.

198. KOEVOETS Wim (12/11/2004)
Essays on Unions, Wages and Employment. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2004.

199. CALLENS Marc (22/11/2004)
Essays on multilevel logistic Regression. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2004.

200. RUGGOO Arvind (13/12/2004)
Two stage designs robust to model uncertainty. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2004.

201. HOORELBEKE Dirk (28/01/2005)
Bootstrap and Pivoting Techniques for Testing Multiple Hypotheses. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2005.

202. ROUSSEAU Sandra (17/02/2005)
Selecting Environmental Policy Instruments in the Presence of Incomplete Compiance. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2005.

203. VAN DER MEULEN Sofie (17/02/2005)
Quality of Financial Statements : Impact of the external auditor and applied accounting standards. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2005.

204. DIMOVA Ralitza (21/02/2005)
Winners and Losers during Structural Reform and Crisis : the Bulgarian Labour Market Perspective. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2005.

205. DARKIEWICZ Grzegorz (28/02/2005)
Value-at-risk in Insurance and Finance : the Comonotonicity Approach. Leuven,

K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2005.

206. DE MOOR Lieven (20/05/2005)
The Structure of International Stock Returns : Size, Country and Sector Effects in Capital Asset Pricing. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2005.

207. EVERAERT Greetje (27/06/2005)
Soft Budget Constraints and Trade Policies : The Role of Institutional and External Constraints. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2005.

208. SIMON Steven (06/07/2005)
The Modeling and Valuation of complex Derivatives : The Impact of the Choice of the term structure model. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2005.

209. MOONEN Linda (23/09/2005)
Algorithms for some Graph-Theoretical Optimization Problems. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2005.

210. COUCKE Kristien (21/09/2005)
Firm and industry adjustment under de-industrialisation and globalization of the Belgian economy. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2005.

211. DECAMPS Marc (21/10/2005)
Some actuarial and financial applications of generalized diffusion processes. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2005.

212. KIM Helena (29/11/2005)
Escalation games: an instrument to analyze conflicts. The strategic approach to the bargaining problem. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2005.

213. GERMENJI Etleva (06/01/2006)
Essays on the Economics of Emigration from Albania. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2006.

214. BELIEN Jeroen (18/01/2006)
Exact and Heuristic Methodologies for Scheduling in Hospitals: Problems, Formulations and Algorithms. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2006.

215. JOOSSENS Kristel (10/02/2006)
Robust discriminant analysis. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2006.

216. VRANKEN Liesbet (13/02/2006)
Land markets and production efficiency in transition economies. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2006.

217. VANSTEENKISTE Isabel (22/02/2006)
Essays on non-linear modelling in international macroeconomics. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2006.

218. WUYTS Gunther (31/03/2006)
Essays on the liquidity of financial markets. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2006.

219. DE BLANDER Rembert (28/04/2006)
Essays on endogeneity and parameter heterogeneity in cross-section and panel data. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2006.

220. DE LOECKER Jan (12/05/2006)
Industry dynamics and productivity. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2006.

221. LEMMENS Aurélie (12/05/2006)
Advanced classification and time-series methods in marketing. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2006.

222. VERPOORTEN Marijke (22/05/2006)
Conflict and survival: an analysis of shocks, coping strategies and economic mobility in Rwanda, 1990-2002. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2006.

223. BOSMANS Kristof (26/05/2006)
Measuring economic inequality and inequality aversion. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2006.

224. BRENKERS Randy (29/05/2006)
Policy reform in a market with differentiated products: applications from the car market. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2006.

225. BRUYNEEL Sabrina (02/06/2006)
Self-control depletion: Mechanisms and its effects on consumer behavior. Leuven,

K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2006.

226. FAEMS Dries (09/06/2006)
Collaboration for innovation: Processes of governance and learning. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2006.

227. BRIERS Barbara (28/06/2006)
Countering the scrooge in each of us: on the marketing of cooperative behavior. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2006.

228. ZANONI Patrizia (04/07/2006)
Beyond demography: Essays on diversity in organizations. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2006.

229. VAN DEN ABBEELE Alexandra (11/09/2006)
Management control of interfirm relations: the role of information. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2006.

230. DEWAELHEYNS Nico (18/09/2006)
Essays on internal capital markets, bankruptcy and bankruptcy reform. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2006.

231. RINALDI Laura (19/09/2006)
Essays on card payments and household debt. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2006.

232. DUTORDOIR Marie (22/09/2006)
Determinants and stock price effects of Western European convertible debt offerings: an empirical analysis. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2006.

233. LYKOGIANNI Elissavet (20/09/2006)
Essays on strategic decisions of multinational enterprises: R&D decentralization, technology transfers and modes of foreign entry. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2006.

234. ZOU Jianglei (03/10/2006)
Inter-firm ties, plant networks, and multinational firms: essays on FDI and trade by Japanse firms. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2006.

235. GEYSKENS Kelly (12/10/2006)
The ironic effects of food temptations on self-control performance. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2006.

236. BRUYNSEELS Liesbeth (17/10/2006)
Client strategic actions, going-concern audit opinions and audit reporting errors. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2006.

237. KESSELS Roselinde (23/10/2006)
Optimal designs for the measurement of consumer preferences. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2006.

238. HUTCHINSON John (25/10/2006)
The size distribution and growth of firms in transition countries. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2006.

239. RENDERS Annelies (26/10/2006)
Corporate governance in Europe: The relation with accounting standards choice, performance and benefits of control. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2006.

240. DE WINNE Sophie (30/10/2006)
Exploring terra incognita: human resource management and firm performance in small and medium-sized businesses. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2006.

241. KADITI Eleni (10/11/2006)
Foreign direct investments in transition economies. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2006.

242. ANDRIES Petra (17/11/2006)
Technology-based ventures in emerging industries: the quest for a viable business model. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2006.

243. BOUTE Robert (04/12/2006)
The impact of replenishment rules with endogenous lead times on supply chain performance. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2006.

244. MAES Johan (20/12/2006)
Corporate entrepreneurship: an integrative analysis of a resource-based model. Evidence from Flemish enterprises. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2006.

245. GOOSSENS Dries (20/12/2006)
Exact methods for combinatorial auctions. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2006.

246. GOETHALS Frank (22/12/2006)
Classifying and assessing extended enterprise integration approaches. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2006.

247. VAN DE VONDER Stijn (22/12/2006)
Proactive-reactive procedures for robust project scheduling. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2006.

248. SAVEYN Bert (27/02/2007)
Environmental policy in a federal state. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2007.

249. CLEEREN Kathleen (13/03/2007)
Essays on competitive structure and product-harm crises. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2007.

250. THUYSBAERT Bram (27/04/2007)
Econometric essays on the measurement of poverty. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2007.

251. DE BACKER Manu (07/05/2007)
The use of Petri net theory for business process verification. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2007.

252. MILLET Kobe (15/05/2007)
Prenatal testosterone, personality, and economic behavior. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2007.

253. HUYSMANS Johan (13/06/2007)
Comprehensible predictive models: New methods and insights. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2007.

254. FRANCKEN Nathalie (26/06/2007)
Mass Media, Government Policies and Economic Development: Evidence from Madagascar. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2007.

255. SCHOUBBEN Frederiek (02/07/2007)
The impact of a stock listing on the determinants of firm performance and investment policy. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2007.

256. DELHAYE Eef (04/07/2007)
Economic analysis of traffic safety. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2007.

257. VAN ACHTER Mark (06/07/2007)
Essays on the market microstructure of financial markets. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2007.

258. GOUKENS Caroline (20/08/2007)
Desire for variety: understanding consumers preferences for variety seeking. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2007.

259. KELCHTERMANS Stijn (12/09/2007)
In pursuit of excellence: essays on the organization of higher education and research. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2007.

260. HUSSINGER Katrin (14/09/2007)
Essays on internationalization, innovation and firm performance. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2007.

261. CUMPS Bjorn (04/10/2007)
Business-ICT alignment and determinants. Leuven, K.U.Leuven, Faculteit Economische en Toegepaste Economische Wetenschappen, 2007.

262. LYRIO Marco (02/11/2007)
Modeling the yield curve with macro factors. Leuven, K.U.Leuven, Faculteit Economie en Bedrijfswetenschappen, 2007.

263. VANPEE Rosanne (16/11/2007)
Home bias and the implicit costs of investing abroad. Leuven, K.U.Leuven, Faculteit Economie en Bedrijfswetenschappen, 2007.

264. LAMBRECHTS Olivier (27/11/2007)
Robust project scheduling subject to resource breakdowns. Leuven, K.U.Leuven, Faculteit Economie en Bedrijfswetenschappen, 2007.

265. DE ROCK Bram (03/12/2007)
Collective choice behaviour: non parametric characterization. Leuven, K.U.Leuven, Faculteit Economie en Bedrijfswetenschappen, 2007.

266. MARTENS David (08/01/2008)
Building acceptable classification models for financial engineering applications. Leuven, K.U.Leuven, Faculteit Economie en Bedrijfswetenschappen, 2008.

267. VAN KERCKHOVEN Johan (17/01/2008)
Predictive modelling: variable selection and classification efficiencies. Leuven, K.U.Leuven, Faculteit Economie en Bedrijfswetenschappen, 2008.